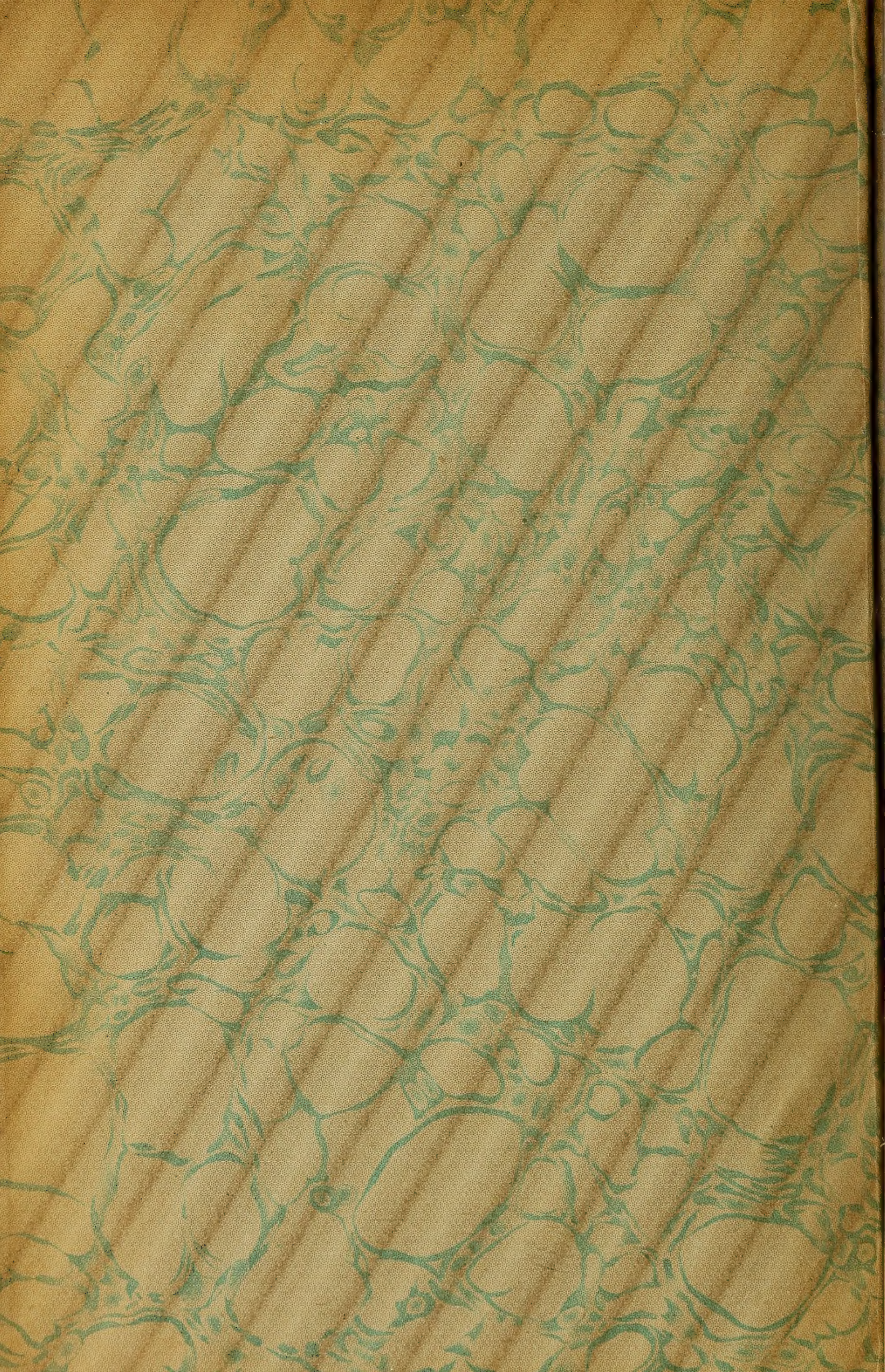
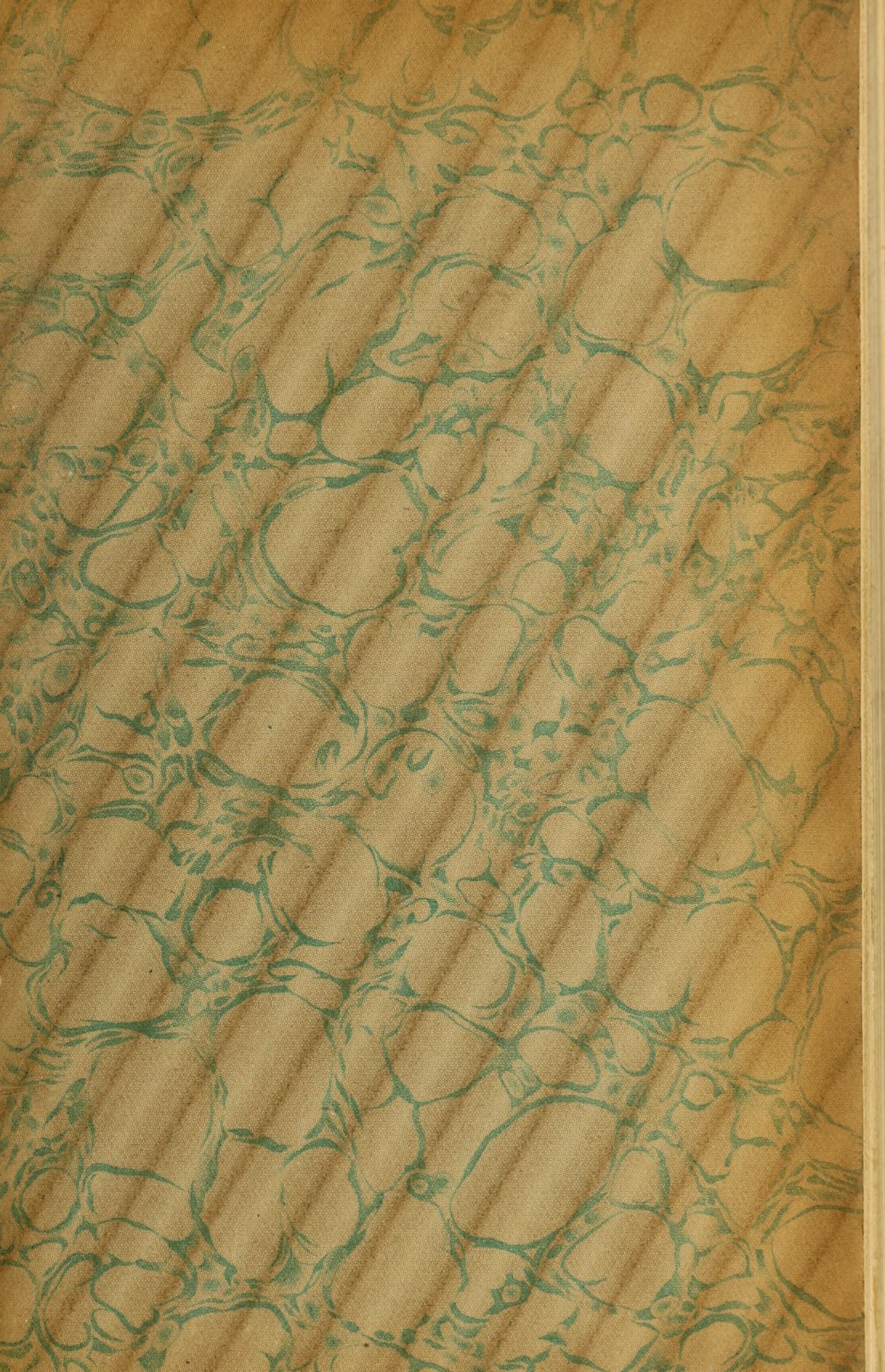




PRINCIPES DE STATISTIQUE

THÉORIQUE ET APPLIQUÉE

PRINCIPES

DE

STATISTIQUE THÉORIQUE

ET APPLIQUÉE

PAR

Armand JULIN

Secrétaire Général du Ministère de l'Industrie, du Travail et du Ravitaillement

Chargé de Cours à l'Université de Gand

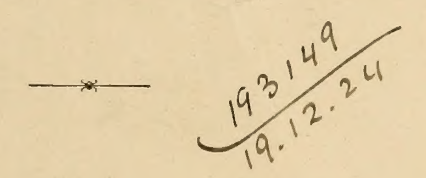
Avec une PRÉFACE DE M. LUCIEN MARCH

Directeur de la Statistique Générale de la France

45 FIGURES & DIAGRAMMES

TOME PREMIER

STATISTIQUE THÉORIQUE



PARIS
MARCEL RIVIÈRE, ÉDITEUR
31, rue Jacob et 1, rue St-Benoît

BRUXELLES
ALBERT DEWIT, ÉDITEUR
53, rue Royale

1921

A MA FEMME.

A MES ENFANTS.

PRÉFACE

L'ouvrage dont la publication commence avec le présent volume est le développement d'un « précis » écrit il y a une douzaine d'années. Il comprend la matière des leçons professées par l'auteur à l'Université de Louvain d'abord, à l'Institut supérieur de commerce d'Anvers ensuite, et vivifiées par les réflexions de l'administrateur qui a appliqué les principes dont il élargit aujourd'hui l'exposé. Mûri au milieu des dures épreuves de la guerre, dans un pays envahi, le premier volume paraît au moment où la nécessité de reconstituer les Etats politiques, de stimuler les forces économiques et les rouages sociaux de ces Etats, rend plus utile que jamais l'observation précise et méthodique des collectivités.

Exercé depuis longtemps aux enquêtes et à l'étude monographique des faits sociaux, M. Julin ne sacrifie point la description analytique des faits au besoin de synthèse, mais il arme cette synthèse de l'appareil mathématique qui en assure la solidité, sans d'ailleurs aller au delà de l'indispensable. Dans le présent volume, l'étudiant trouvera un exposé simple de méthodes aujourd'hui classiques, illustré par de nombreuses applications numériques qui rendent sensibles l'enchaînement et la portée de ces théories. Les deux autres volumes se rapporteront, l'un à la statistique économique, l'autre à la statistique du travail.

Les trois parties de ce premier volume renferment, la première un aperçu des caractères généraux de la statistique, la seconde l'exposé des procédés communément employés pour le récolement des données numériques que la statistique met en œuvre, la troisième l'indication de la méthode qui préside à cette mise en œuvre.

La dernière partie est celle qui assigne à la statistique sa valeur propre : les matériaux ne servent réellement que lorsque la construction est achevée; la construction n'est solide que si ses parties sont convenablement ordonnées et ajustées avec une certaine précision. Le langage mathématique facilite singulièrement l'exposition des règles qui permettent d'atteindre ce résultat; l'auteur l'emploie avec sobriété, cherchant surtout à faire saisir par des exemples le mécanisme et l'application des formules.

Les principes qui interviennent dans le traitement des observations statistiques ne diffèrent pas de ceux que l'on applique dans les sciences de la nature, toutes les fois que l'on recherche une certaine précision en vue de serrer la réalité de plus près. Toute science ne se réduit-elle pas à un système dans lequel on fait état d'analogies ou de ressemblances, et de différences ou de dissemblances, pour former des catégories uniformes et faciliter le travail de la pensée? Moins ces catégories sont nombreuses, plus la pensée évolue avec aisance; par contre plus elle s'éloigne de la réalité, plus elle perd l'acuité qui lui faisait embrasser, d'une façon plus ou moins confuse, les nuances infinies des qualités des choses. Les expériences célèbres de Regnault ont provoqué l'indignation de certains savants : elles ont pourtant préparé des progrès considérables; par contre, aucune collection de mesures, si nombreuses soient-elles, ne nous donne une image des mouvements de la mer aussi exacte que la vue directe, réfléchie par la pensée, ou que l'audition du poème symphonique qui nous place dans le même état d'âme.

Mais le développement de la science exige des efforts : ceux auxquels la pensée est condamnée par la multiplication de catégories qui sont de plus en plus étroitement délimitées, grâce à la précision avec laquelle on note les ressemblances et les différences. Accroître cette précision, c'est permettre à la science de tendre vers la finesse de la perception confuse tout en conservant la clarté, la fixité, l'uniformité des notions conventionnelles. Parfois d'ailleurs cette précision fait apparaître des analogies insoupçonnées et merveilleusement fécondes, quand des mesures équivalentes se rapportent à des phénomènes qui semblaient d'essence très différente.

Dans tous les cas, la précision croissante des mesures indique avec évidence que l'uniformité complète n'existe que dans les notions conventionnelles. On saisit mieux alors la nécessité pratique d'une règle telle que celle de la moyenne et des lois qui résultent d'interpolations.

La notion de précision est d'ailleurs toute relative, variant avec le progrès des instruments. Dans les sciences de la nature inanimée, où le degré de précision dont on se contentait il y a un siècle a été centuplé, on est à même de prévoir certains événements avec une extrême rigueur. Dans les sciences de la vie, dans les sciences sociales surtout, où les choses se mesurent à l'échelle de l'homme et non à l'échelle des éléments de la matière, on se contente d'approximations relativement larges, suffisantes pour la vie pratique.

Ce n'est pas seulement pour prévoir, c'est aussi pour raisonner juste qu'une certaine précision est nécessaire, car les notions issues de la réalité n'en épousent pas exactement la forme: il importe de limiter la déformation par un criterium universellement adopté.

De là l'utilité du langage mathématique, — qui d'ailleurs peut souvent être réduit à des expressions simples —, toutes

les fois que l'on raisonne sur des grandeurs, même quand ces grandeurs se rapportent à la vie de l'homme en société.

Il importe seulement de ne point oublier que l'étude numérique, de même que l'étude analytique, des faits observés ne peut donner plus que le contenu de ces faits. Aussi convient-il de passer les observations au crible d'une critique attentive. « Le développement de la critique, remarque M. Julin, a toujours marché de pair avec l'esprit scientifique. » On pourrait dire encore que l'esprit critique est la première condition de l'esprit scientifique : on ne construit bien que si l'on est à même d'apprécier la valeur des matériaux. Le véritable idéal du statisticien est l'application de l'esprit scientifique à la mise en œuvre des observations numériques. Imprégné de cet idéal, le statisticien ne s'embarrassera guère de la question de savoir si la statistique est une méthode, une science, un groupe de sciences ou un art, ou bien si la statistique peut légitimement intervenir dans de nombreux ordres de connaissances, si au contraire elle doit limiter son domaine à la mesure des phénomènes sociaux.

Les controverses à propos de classification naissent généralement d'un certain désaccord sur les limites assignées aux catégories dont on discute. On pourrait aussi se demander si la géométrie est une méthode ou une science; on pourrait comparer l'unité de la physique d'Aristote à l'unité d'une discipline particulière telle que la morphologie des plantes, la géochimie ou la mécanique statistique. Il importe moins de savoir si l'étude à laquelle on consacre son activité est ou non une science, que de poursuivre cette étude avec un état d'esprit scientifique, car c'est le seul moyen de déterminer l'acceptation universelle des résultats.

La nécessité de n'aborder les recherches statistiques qu'avec cet état d'esprit est surtout évidente quand il s'agit de statistiques internationales où se heurtent

souvent des tendances respectables mais opposées. Que de sécurité dans l'état économique des peuples, dans leurs relations permanentes, si l'on connaissait à tout moment, avec quelque précision, d'une part l'état des ressources en produits consommables, quantités produites et stocks, en rapport avec les besoins de la consommation, d'autre part la puissance des moyens de production et des éléments de progrès dans chacun d'eux. Mais ce ne sont point des données qu'il soit possible de grouper incontinent par la vertu magique d'une enquête accidentelle. Il faut une observation patiente et de longue durée pour que la documentation statistique, arrivant, par des milliers de canaux, des principaux points du globe, se résume en une matière clarifiée et de bon aloi.

De même que les langues diffèrent d'une contrée à l'autre, souvent d'une région à l'autre d'un même pays, le sens et la compréhension des termes qui servent à désigner les catégories, dans les tableaux statistiques, varient suivant les pays, suivant les régions.

C'est ainsi que dans les comparaisons internationales, les rubriques qui caractérisent les âges des individus, les enfants nés vivants ou morts, les religions, les crimes et délits, les degrés d'instruction, les professions exercées, les grèves, les marchandises échangées, les unités de transport, les qualifications d'origine, la circulation fiduciaire, les charges financières publiques, etc., comportent souvent des acceptions différentes; les chiffres correspondant à une même rubrique s'appliquent à des choses qui ne sont point les mêmes. La principale tâche des organismes internationaux tels que l'Institut international de statistique est d'obtenir une plus grande uniformité, de tendre vers l'unité des relevés, des cadres et des notations statistiques. Mais les difficultés sont grandes; de longs et patients efforts sont nécessaires; longtemps encore on ne pourra légitimer, sur beau-

coup de points, les comparaisons internationales qu'en les accompagnant de commentaires détaillés c'est-à-dire d'un examen critique attentif des résultats comparés.

On ne saurait d'ailleurs renoncer à ces comparaisons, sous le prétexte qu'elles sont incertaines et dangereuses. Si la vie internationale peut comporter un foyer de lumières communes, un centre de relations permanentes, il importe que la précision et la rapidité des informations sur tout ce qui intéresse l'existence des peuples, s'améliorent sans cesse. Seule la pratique de ces informations peut apprendre dans quel sens et par quels moyens doit se réaliser le mieux. On n'aperçoit en effet l'insuffisance des relevés numériques que lorsqu'on en analyse les éléments, que l'on sonde les différences constatées. Il ne faut point sans cesse fuir le danger, puisqu'il naît de l'intensité de la vie; il importe seulement d'y parer.

Loin de proscrire les rapprochements douteux, il convient au contraire de les encourager, à la condition que l'on accompagne les chiffres comparés des éclaircissements et des explications nécessaires pour éviter les erreurs d'interprétation; il n'est guère d'autre moyen d'améliorer peu à peu l'élaboration des statistiques, de mettre en relief les défectuosités et d'atteindre un jour l'uniformité désirable des relevés.

Après le bouleversement qui a mis tant de peuples aux prises, mais qui en a rapproché aussi et qui a créé des liens nouveaux, le désir que l'on ressent de toute part de voir ces liens se fortifier, conduit à multiplier et à rendre plus intimes les relations internationales et sociales, à les sceller de cette confiance mutuelle qui exige une connaissance suffisamment étendue des circonstances de la vie économique et de la vie sociale dans le monde nouveau. Coordonner les informations en évitant les doubles emplois, passer ces informations au crible d'une critique judicieuse, mettre au point les écarts et les rapports constatés: telle

est l'œuvre scientifique qui aidera les nations à se mieux connaître et à apprécier leur valeur respective dans l'effort commun vers la prospérité générale et la justice.

Pour réaliser cette œuvre ou pour en détacher les fruits les plus utiles, des ouvriers instruits, exercés au travail scientifique, pénétrés des principes de la méthode statistique, sont nécessaires. L'ouvrage que nous présentons aujourd'hui au lecteur apprend à bien observer les faits dont s'occupe la statistique, à apprécier la valeur des observations et à les mettre en œuvre pour en tirer les enseignements les plus sûrs: c'est un excellent guide où chacun puisera ce qui est essentiel dans la théorie et dans la pratique du travail statistique.

LUCIEN MARCH.

AVERTISSEMENT

La division de l'ouvrage, sa genèse, les circonstances dans lesquelles il a été écrit sont indiquées dans la Préface ; nous n'avons rien à y ajouter. Il est utile toutefois d'insister sur ce point que nous avons voulu donner à notre travail l'armature mathématique indispensable, mais en évitant de rebuter le lecteur et de tomber dans l'exagération qui semble faire de la statistique une simple division des mathématiques. Sans une bonne observation, il n'y a pas de bonne statistique ; sans une critique serrée, il n'y a pas d'esprit scientifique ; sans esprit scientifique, il n'y a pas d'utilisation rationnelle des matériaux : ce sont ces quelques idées simples qui nous ont constamment guidé.

Au cours de ce long travail, poursuivi dans des circonstances pénibles, nous avons eu le réconfort de précieux encouragements. Nos remercîments vont surtout à nos amis : notre collègue, M. Ch. De Lannoy, professeur à l'Université de Gand ; M. Beaujean, directeur de la Caisse d'Épargne et de Retraite sous la garantie de l'État, à Bruxelles, et M. Lucien March, directeur de la Statistique Générale de la France, qui, avec un soin dont nous leur savons infiniment gré, ont revu le manuscrit. Nous leur sommes redevable de maintes améliorations et corrections ; nous leur exprimons ici toute notre reconnaissance.

S'intéressant dès le début à notre œuvre, notre ami, M. F. Carpentier, a bien voulu prendre sa part de travail en contrôlant de nombreuses opérations numériques et en réunis-

sant les notes d'après lesquelles ont été rédigés les nos 321-324 et 336-338; qu'il trouve ici l'expression de nos remercîments cordiaux.

Ce volume comprend de très nombreuses applications numériques relatives surtout à la statistique économique et sociale; l'auteur espère que le soin apporté aux calculs les aura préservées de toute erreur; il serait reconnaissant au lecteur de lui signaler celles qui auraient pu lui échapper.

A. J.

Bruxelles, le 15 novembre 1920.

TABLE DES MATIÈRES [1].

(Les chiffres en caractères romains renvoient aux chapitres ou paragraphes ; ceux en chiffres ordinaires aux numéros.)

INTRODUCTION.

CHAPITRE PREMIER.

Phénomènes étudiés par la statistique.

CHAPITRE II.

Différentes conceptions de la statistique.

CHAPITRE III.

Caractères propres à la statistique.

CHAPITRE IV.

---

(1) Voyez à la fin du volume : A, la table analytique; B, l'index onomastique.

*

LIVRE PREMIER.

Technique du relevé statistique.

SECTION I. — Le relevé statistique ou relevé direct.

CHAPITRE PREMIER.

Généralités, définition, divisions.

CHAPITRE II.

Organisation du relevé statistique.

SECTION II. — Le relevé indirect.

CHAPITRE PREMIER.

Généralités, définition, divisions.

CHAPITRE II.

L'enquête et la monographie.

SECTION III. — La critique statistique.

CHAPITRE PREMIER.

Généralités, définitions, divisions.

CHAPITRE II.

Critique de sincérité.

CHAPITRE III.

Critique d'exactitude.

CHAPITRE IV.

SECTION IV. — Le dépouillement et la présentation des données statistiques.

CHAPITRE PREMIER.

Préparation du dépouillement.

CHAPITRE II.

Exécution du dépouillement.

CHAPITRE III

Le dépouillement statistique et le calcul par les machines

CHAPITRE IV.

La présentation des résultats statistiques.

LIVRE II.

Procédés d'analyse du matériel statistique.

CHAPITRE PREMIER

Séries, sériation, distribution.

CHAPITRE II.

Moyennes, médiane, dominante.

CHAPITRE III.

La dispersion et ses mesures.

CHAPITRE IV.

Covariation (*Corrélation*).

CHAPITRE V.

Statistique graphique.

A. — Diagrammes.

B. — Cartogrammes.

C. — Stéréogrammes.

LIVRE III.

La loi des erreurs.

INTRODUCTION

CHAPITRE PREMIER

Phénomènes étudiés par la statistique.

I. — Phénomènes typiques et phénomènes collectifs.

1. A les considérer sous l'aspect de la méthode applicable à leur étude, les phénomènes se partagent en deux catégories. Les uns sont strictement typiques : chaque cas individuel est identique à un autre cas de même nature, se présentant dans des conditions semblables. Des expériences répétées aboutissent partout au même résultat. Que l'on procède à ces expériences, dans des conditions diverses de temps et de lieu, on verra que la règle de formation du phénomène reste une. Les sciences appliquées à la nature inorganique fournissent les exemples les meilleurs dans cet ordre d'idées. La réaction chimique de deux corps est constante : si l'on remarque une déviation par rapport aux résultats précédents, c'est une preuve que les conditions de l'expérience n'ont pas été identiques. Une dissolution de soufre dans du sulfure de carbone, abandonnée à froid à l'évaporation spontanée, cristallise en cristaux octaédriques : chauffés à 110°, ces cristaux se transforment en un agrégat de cristaux prismatiques. La millième expé-

rience ne diffère pas de la première et nous atteignons à la connaissance de la norme par une seule expérimentation.

Les autres phénomènes, au contraire, qu'on appelle phénomènes collectifs, sont atypiques; leurs manifestations varient sans qu'on puisse leur attribuer une règle fixe. Ils sont extrêmement nombreux. On les rencontre non seulement parmi les faits sociaux, mais aussi dans les sciences zoologiques ou botaniques, en météorologie comme en médecine ou en économie politique. Lorsqu'on étudie la composition d'un groupe de population, on se trouve en présence de phénomènes collectifs, mais la démographie n'est pas l'unique domaine où ils se manifestent en aussi grand nombre, ils sont tout aussi fréquents en météorologie, par exemple. Ceci est important sous le rapport méthodologique et nous aurons à y insister plus tard. Ce que nous retenons à présent, c'est que les phénomènes collectifs, essentiellement différents de ceux qu'on appelle typiques, ne se rencontrent pas exclusivement parmi les faits sociaux, ainsi que la réflexion l'indique à chacun sans difficulté.

2. Examinons brièvement deux exemples de phénomènes collectifs, sous le rapport de la méthode d'observation qui s'impose en ce qui les concerne.

La classe ouvrière est exposée à un grand nombre d'accidents résultant du travail. Ce nombre n'est pas le même chaque année; les accidents ne sont pas aussi fréquents dans toutes les industries; leur gravité varie, comme aussi les charges que leur réparation entraîne, la nature des lésions survenues, la durée et l'étendue de l'incapacité de travail qui en résulte, etc. Il ne suffit donc pas d'observer quelques accidents de travail pour être informé de la gravité du mal; il faut les compter tous, les observer chacun en particulier, les comparer au nombre d'ouvriers exposés à un même risque, en connaître la gravité et les conséquences, enfin cataloguer ces cas multiples d'après des catégories qu'à première vue nous pouvons imaginer fort

nombreuses et compliquées. Alors seulement nous pourrons nous former une notion de l'accident du travail, car les manifestations de ce fait varient à l'infini et nous ne pouvons savoir à l'avance à quelles lois elles obéissent. Ce n'est pas en observant une seule ou plusieurs de ces manifestations que nous y parviendrons, mais en les relevant toutes et en tenant compte des conditions dans lesquelles elles se produisent.

Sous nos climats occidentaux, il pleut souvent et les heures ensoleillées nous sont parcimonieusement mesurées. Pour savoir dans quelle mesure la pluie tombe, ou quelle est la fréquence de l'apparition du soleil, nous ne pouvons nous contenter de noter qu'il pleut ou qu'il ne pleut pas tel jour, que le ciel est resté couvert ou que le soleil a brillé. Il nous faut, au contraire, nous livrer à des observations régulières pendant une série d'années, répartir les observations d'après les mois pendant lesquels elles sont faites, mesurer la quantité d'eau tombée, compter les heures de soleil, etc. D'après ces calculs, nous serons en mesure de définir, à ce point de vue particulier, le climat du pays que nous habitons; encore ne pourrons-nous le faire avec certitude que pour les régions qui avoisinent les postes d'observation. Le climat est un phénomène collectif, dont les manifestations sont variables, et dont la norme ne peut se définir qu'à l'aide d'un grand nombre d'observations.

La méthode qui observe les variables pour en déduire le normal ou le typique convient essentiellement aux phénomènes collectifs, comme l'expérimentation convient aux phénomènes typiques : c'est la méthode statistique.

3. La différence profonde qui sépare les phénomènes typiques des phénomènes collectifs, en ce qui concerne le mode de leurs manifestations et, par conséquent, la méthode d'après laquelle il convient de les observer, tient évidemment à la simplicité, d'une part, à la complexité, d'autre part, des causes agissantes. Dans la nature inorganique, les

causes qui interviennent sont constantes; pourvu que les conditions dans lesquelles elles opèrent soient semblables, elles reproduisent mécaniquement les mêmes effets. Elles ne pourraient en produire d'autres. L'expérimentation réalisée dans des conditions rigoureuses ne peut donc que répéter des résultats identiques; aussi est-il inutile de la multiplier. Dans les faits collectifs, au contraire, il y a une multitude de causes opérantes qui s'entre-croisent, s'enchevêtrent et produisent des différenciations infinies dans les manifestations des phénomènes. D'où cette conséquence que si l'on veut dégager les caractères les plus fréquents, les tendances les plus communes, il faut nécessairement observer un grand nombre d'unités, afin de réunir tous les cas possibles et classer ensemble ceux qui présentent des analogies. Cette besogne de classement est d'autant plus longue et présente d'autant plus de difficultés que les causes agissantes sont nombreuses et enchevêtrées. Mais il n'y a pas d'autre moyen de parvenir à la connaissance des faits collectifs et l'on peut dire en toute raison que le procédé statistique est indispensable pour décrire et analyser les groupes trop étendus ou trop complexes qui ne peuvent être saisis par la simple observation scientifique isolée (1).

4. Si les phénomènes collectifs étaient soumis à des variations incessantes, telles que les traits observés aujourd'hui ne seraient plus vrais demain, il n'y aurait qu'un intérêt fugitif à les observer; on pourrait même dire que cet intérêt serait trop mince pour justifier la peine et la dépense que coûtent les observations portant sur des masses considérables. C'est l'opinion que se forment souvent de la statistique certaines personnes qui n'ont qu'une notion incomplète du concept de loi sociale, économique et statistique. Parce qu'elles se rendent compte de ce que le monde éco-

(1) Wagner, *Les fondements de l'économie politique*, t. I, p. 289. Paris, Giard et Brière, 1904.

nomique, par exemple, est soumis à une évolution incessante, elles se persuadent facilement que les recherches de la statistique sont oiseuses. A quoi bon s'efforcer de saisir une situation changeante à tout moment, et de quelle utilité peuvent être des constatations qui se trouvent depuis longtemps périmées lorsqu'elles sont livrées à la publicité? Cette opinion, assez répandue, ne tient pas compte de la nature des causes qui interviennent dans les phénomènes collectifs. Ces causes sont de deux espèces. Les unes ont quelque chose de général et de permanent. Elles sont communes à toutes les manifestations individuelles du phénomène. Parfois, elles se manifestent d'une façon continue; parfois, elles ne font sentir leurs effets qu'à un moment donné, mais d'une façon néanmoins régulière (1). D'autres causes, au contraire, présentent le caractère opposé. Elles sont propres à une situation donnée, elles n'intéressent qu'une manifestation isolée d'un phénomène, elles n'exercent pas d'action sur toutes les manifestations, mais seulement sur quelques-unes et en nombre variable. On peut donc dire qu'au lieu d'être générales, elles sont spéciales, et accidentelles au lieu d'être permanentes. Dans les phénomènes collectifs, les deux catégories de causes se retrouvent, elles s'entre-croisent, elles se contrarient, elles s'ajoutent ou se contre-balancent. Toutefois, il est évident que de ces deux sortes de causes, celles qui ont un caractère général et permanent exerceront une action beaucoup plus décisive que les autres.

Les causes accidentelles ont une tendance à se faire équilibre, à s'annihiler réciproquement. Le but de la recherche s'exerçant sur les phénomènes collectifs est de dégager les

(1) Ce sont les causes que Quetelet appelait causes *constantes*, à raison de la continuité de leurs manifestations et de l'égalité de leur intensité, comme il nommait causes *variables*, celles dont l'action continue se présentait avec une énergie sujette à changement. Knapp avait proposé de réunir sous le nom de *causes essentielles* les causes constantes et les causes variables. (Cf. J. Lottin, « Le calcul des probabilités et les régularités statistiques », *Revue Néo-Scolastique*, 1910, pp. 45-46).

causes agissantes diverses et particulièrement les causes générales, de façon à exprimer ce que le phénomène présente de permanent et de typique, comme aussi de montrer les modifications que les éléments permanents subissent au cours des années, ce qui permet de tracer la courbe d'évolution du phénomène. Ainsi, l'on peut dire que la statistique, méthode propre aux phénomènes collectifs, a pour objet final la recherche de l'absolu parmi le relatif, du typique parmi l'accidentel, du permanent parmi le passager.

II. — Notation numérique des observations.

5. La nécessité de consigner les résultats de l'observation sous forme de données numériques est l'un des caractères fondamentaux du procédé statistique; on peut la placer immédiatement après la distinction entre les phénomènes typiques et les phénomènes collectifs. Non seulement elle exerce une influence décisive sur toute la technique, mais elle a aussi une action prépondérante sur les méthodes critiques et les procédés d'interprétation à employer. C'est à tort que certains auteurs ont semblé n'attacher qu'une importance secondaire à ce caractère de la statistique et ne lui ont accordé que peu d'attention. C'est une erreur plus grande encore de croire que, dans certains cas, la statistique peut se passer de la notation numérique. D'après Wagner, les situations créées par les résistances des sujets de l'observation, leur état de civilisation, leurs préjugés pouvant mettre obstacle à la recherche statistique, il serait alors permis de se contenter d'évaluations approximatives, faute de nombre précis. Les expressions générales, dit-il, telles que : beaucoup, peu, plus, moins, plus grand, moindre, etc., ne doivent pas être bannies de la méthode statistique lorsque les données plus exactes font défaut.

Wagner reconnaît cependant que dans les cas qu'on vient d'indiquer, la méthode statistique perd de sa valeur. C'est

en condamner l'emploi dans les cas que l'on vise, car on ne conçoit pas une recherche scientifique basée sur deux méthodes : l'une qui est reconnue bonne, l'autre qui est certainement défectueuse. Cette dernière n'est pas à employer, et si l'on n'a pas la possibilité de recourir à la bonne méthode, la seule existante sous le rapport scientifique, il est bien préférable de conclure que la recherche, dans les conditions actuelles, n'est pas possible.

Le chimiste, le physicien, le biologiste emploient indifféremment l'observation et l'expérimentation, afin d'arriver à caractériser les phénomènes qu'ils étudient. Comme ce sont des phénomènes typiques, l'observateur n'a pas à se préoccuper de l'aspect quantitatif; il n'a qu'à envisager l'aspect qualitatif, ce qui fait que ses remarques prennent naturellement la forme écrite descriptive.

Le statisticien, au contraire, étudie des phénomènes collectifs atypiques : pour les décrire, il recourt à l'énumération, au dénombrement, il les considère sous le rapport quantitatif. C'est en chiffres qu'il traduit ses observations.

La notation numérique est un procédé propre à la méthode statistique, qui ne se rencontre pas dans les autres et qui est le seul dont on puisse se servir à propos des phénomènes collectifs pour la connaissance desquels la méthode statistique est indispensable. A raison de son importance, il ne sera pas inutile de s'y arrêter quelque peu.

6. Lorsque le relevé est achevé et que les divisions de la classification sont établies, le statisticien, au moyen du dépouillement, compte le nombre de fois que se rencontre telle variété appartenant à un groupe donné de classement, par exemple le nombre de salaires de 3 francs à fr. 3.49 existant dans un groupe ouvrier dont il étudie la distribution sous le rapport du salaire. Les chiffres inscrits par le statisticien dans ses tableaux signifient que l'unité dont il fait usage se rencontre autant de fois sous l'aspect de telle variété : les chiffres expriment donc une idée de *fré-*

quence qui sert à interpréter le phénomène collectif dont les manifestations individuelles sont variables. Ainsi, par exemple, les établissements industriels d'un pays n'ont pas tous la même importance. En admettant que le nombre d'ouvriers occupés dans chaque établissement soit un indice suffisant, on commence par isoler ce caractère pour chaque entreprise, puis on partage la série en un certain nombre de classes : établissements comptant de 1 à 5 ouvriers, de 6 à 10, de 11 à 20, etc., et dans chaque colonne on inscrit le nombre d'entreprises qui présentent ce caractère. Cette statistique indique que sur la masse des entreprises, celles qui occupent de 1 à 5 ouvriers, etc., se rencontrent autant de fois. De là, on déduira ensuite que la petite, la moyenne, la grande industrie, sont plus ou moins représentées dans l'ensemble. On fera varier ce classement en l'établissant par industrie et en recherchant quelles sont les différences qui se marquent dans la grandeur des établissements, selon qu'ils appartiennent à telle industrie ou à telle autre. On étendra encore cette recherche en ajoutant la base de l'unité territoriale et en étudiant le rapport qui se marque entre la grandeur des entreprises et leur répartition géographique, en d'autres termes, en recherchant s'il existe des zones de grande et de petite industrie, etc. On procède donc d'après les règles de la connaissance intellectuelle, par abstraction, c'est-à-dire que l'on considère successivement chaque caractère après un autre caractère, procédé que rend seule possible la notation numérique des observations.

7. La notation numérique présente des avantages sérieux, comme aussi elle a ses dangers et ses inconvénients. Au nombre de ses avantages, on doit lui reconnaître la précision. Il est certain que la présentation, sous forme de tableaux, d'une série bien ordonnée de chiffres, recueillis à l'aide d'une technique parfaite, doit aider beaucoup à la découverte des causes et des tendances, c'est-à-dire contri-

buer au but de la statistique. La disposition tabellaire des données de la statistique a toujours été considérée comme un perfectionnement sur les procédés antérieurement en usage, à l'époque où la statistique n'était encore considérée que comme une science descriptive de l'Etat : « Les tableaux, dit Engel, peuvent se comparer à un ensemble de fonctions de diverses natures; les valeurs de la première colonne correspondent aux données indépendantes, aux constantes; celles-ci une fois bien établies, les nombres des colonnes suivantes varient avec les premières, et sont comme les variables dépendantes des constantes. »

Ainsi, dans une statistique des salaires, les variables sont constituées par les nombres d'ouvriers gagnant chaque taux de salaire, les constantes étant représentées par la nature de l'industrie, ou une base géographique, ou le caractère de grande ou de petite industrie. Les différents tableaux qu'on obtient de la sorte font connaître l'influence que la nature de l'industrie, ou la localisation de l'entreprise, ou la grandeur des établissements peuvent exercer sur la répartition des salaires parmi une population ouvrière donnée.

Le fait que les observations statistiques se traduisent en chiffres est la cause de certains inconvénients. Au premier rang de ceux-ci nous notons la trop grande rigidité apparente de certaines constatations. Parce qu'elles sont exprimées en chiffres, les conclusions de la statistique paraissent empreintes d'une certitude absolue. On ne discute pas contre un chiffre, dit-on; non, mais on discutera utilement la méthode à l'aide de laquelle ce chiffre a été obtenu, et on pourra ainsi discuter sa signification véritable. Il faut éviter avec soin que l'esprit critique ne soit annihilé par l'appareil rigide dont sont entourées les données statistiques.

8. La notation numérique consiste essentiellement en un relevé quantitatif, mais il serait faux de considérer que la

statistique n'envisage que le côté quantitatif des phénomènes qu'elle observe. Elle est également basée sur l'aspect qualitatif des choses ; seulement les deux aspects ne sont pas envisagés en même temps, ou dans la même phase des opérations techniques. Dans un relevé démographique portant sur une population classée d'après l'état civil, on divisera cette population en plusieurs classes : célibataires, mariés, veufs, divorcés ou séparés. L'opération de classement par laquelle on a reconnu qu'il y avait lieu de diviser la population en quatre classes quand on désire l'étudier sous le rapport de l'état civil des personnes qui la composent, cette opération répond à une distinction qualitative et elle s'effectue du moment où le statisticien arrête les classifications qui vont lui servir. Cette classification, que nous venons de citer, est très simple, mais il en est de plus complexes et dont les éléments, moins clairement indiqués par la nature des choses, peuvent être discutés. Dans une statistique industrielle, nous voulons, par exemple, répartir les entreprises industrielles d'après le mode d'exploitation. Nous pouvons concevoir des distinctions basées sur les principes suivants de différenciation : 1° les entreprises peuvent être des entreprises individuelles ou des associations de personnes ou des associations de capitaux ; 2° dans les entreprises individuelles ou qui sont constituées sur la base d'associations de personnes, on peut reconnaître des entreprises de minime importance dans lesquelles l'exploitant travaille seul, ou bien dans lesquelles il travaille avec des membres de sa famille seulement ; à un degré supérieur, on reconnaîtra les entreprises qui ont recours à des ouvriers étrangers à la famille de l'exploitant, etc. Le travail de classification est par sa nature qualitatif. C'est lorsque les classifications sont établies que la phase quantitative se déroule dans l'opération statistique. L'opérateur reconnaît alors que l'unité qu'il considère appartient à une classe déterminée, il la note dans cette classe et finalement il compte le nombre de fois que des unités de cette sorte se

sont rencontrées dans le relevé. Cette opération est la suite de la première, qui consiste à arrêter une place au programme de classement d'après lequel on comptera les unités dénombrées.

La statistique présente donc dans ses opérations diverses le caractère des méthodes scientifiques : elle abstrait, elle considère les choses selon le point de vue qualitatif et quantitatif, elle aborde successivement l'aspect discriminatif et unitif des phénomènes qu'elle considère.

CHAPITRE II

Différentes conceptions de la statistique.

I. — Développement de la statistique.

9. La conception de la statistique que se sont faite les savants a varié au cours des temps. Pour cette raison, il semble indispensable de faire brièvement un historique de la statistique, afin de montrer comment elle a été comprise et de dégager à la fin la notion vraiment scientifique que l'on peut s'en faire. Cependant, ce travail n'essaye en aucune façon de retracer l'histoire de la statistique. La raison pour laquelle nous nous en abstenons est double : d'abord parce que l'histoire de la statistique a été faite souvent et avec les détails les plus étendus; pour en avoir une idée, il suffit de parcourir les ouvrages de Gabaglio, de Meitzen ou de Wagner, dans lesquels une énorme érudition a été dépensée, à ce point qu'on ne peut guère espérer y ajouter quelque chose d'important. Ensuite parce que ce n'est pas le point de vue historique qui nous intéresse, mais l'aspect méthodologique. Toutefois, dans la longue chaîne d'ouvrages et d'auteurs, il y a lieu de noter quelques

sommets, afin de rattacher à des œuvres connues et à des noms souvent cités en statistique, les théories générales dont la succession sera envisagée au cours du présent exposé.

10. On trouve dans l'antiquité comme au moyen-âge de nombreux exemples de travaux statistiques exécutés sur l'ordre des gouvernements. L'Egypte fit enregistrer tous les chefs de familles, la Judée eut ses recensements de la population, la Chine ses descriptions statistiques du territoire, la Grèce ses dénombrements servant tantôt à répartir les terres entre la population, tantôt à établir l'assiette des impôts, tantôt, enfin, à déterminer le nombre des citoyens et la condition des habitants. Rome fit un fréquent usage des recensements : le census en présence du censeur se répéta 69 fois; on notait, dans des temples désignés, les naissances, la puberté et les décès (1). Le moyen-âge n'ignora pas les travaux de statistique. Le *Domesday book*

---

(1) Le recensement (*censere*) se faisait au Champ de Mars et se répétait tous les cinq ans. Un censeur, tiré au sort, présidait l'assemblée. Le recensement servait de base à l'impôt et s'adressait uniquement aux « assidui » ou « capite censi », opposés aux « proletarii », lesquels étaient placés en dehors des cinq classes basées sur la fortune. Lors du recensement, chaque citoyen devait déclarer ses noms et prénoms, son âge, sa famille, le nombre de ses esclaves et ses biens de toute nature. C'est à l'occasion du cens que les censeurs exerçaient leur pouvoir disciplinaire par la *nota* placée en regard du nom : elle entraînait la suppression du droit de suffrage (*in aerarios referre*).

Le texte suivant d'Aulu-Gelle prouve que le recensement était effectué avec une grande rigueur :

Inter censorias severitates tria haec exempla in litteris sunt castigatissimae disciplinae. Unum est hujuscemodi. Censor agebat de uxoribus solemne jusjurandum. Verba erant ita concepta : « Et tu, ex animi sententia, uxorem habes ? » Qui jurabat, cavillator quidem et canicula et nimis ridicularius fuit. Is, locum esse sibi joci dicendi ratus, quum ita, uti mos erat, censor dixisset : « Et tu, ex animi tui sententia, uxorem habes ? — Habeo equidem, inquit, uxorem, sed non hercle ex animi mei sententia. » Tum censor eum, quod intempestive lascivisset, in aerarios retulit, causamque hanc joci scurrilis apud se dicti subscripsit.

Aulus Gellius, *Noctes Atticae*, liv. IV, chap. XX. Paris, C.-L.-F. Panckoucke, éditeur, 1845, t. I, p. 312.

de Guillaume le Conquérant rentre dans la catégorie des relevés statistiques, comme auparavant certains relevés exécutés sur l'ordre de Charlemagne, comme plus tard de nombreux inventaires, établissement de rôles ou descriptions, dont on peut retrouver des types dans beaucoup de pays de l'Europe centrale et occidentale. Les matières sur lesquelles portèrent ces investigations font partie, incontestablement, du domaine assigné à la statistique, mais il faut renoncer à trouver, dans ces recherches entreprises dès l'antiquité et le moyen âge, aucun principe de technique et aucune idée précise de la nature scientifique de la statistique. Ce sont des vues fragmentaires, empiriques, dominées par les circonstances du moment qui décident de l'opportunité et de la nécessité de tel genre de recherches ; quant à la technique, elle est dans l'enfance et est plus ou moins satisfaisante selon l'application de celui qui la conçoit, la conscience de ceux qui l'appliquent et la bonne volonté de ceux qui doivent se soumettre aux investigations. Il faut que le besoin d'ordre et de connaissances soit devenu général pour que la statistique s'organise et que sa notion exacte se dégage peu à peu de l'empirisme qui a marqué ses débuts.

11. Lorsque l'Etat moderne s'organisa, on sentit de plus en plus le besoin de réunir des informations exactes sur tous les éléments qui le composent et desquels sa prospérité dépend. La population, l'armée, les finances, l'industrie, le commerce sont autant de points sur lesquels les investigations des statisticiens de l'époque se portent avec curiosité. Aujourd'hui encore ces matières sont celles que la statistique aborde le plus communément, mais elle le fait dans un autre esprit. Les auteurs qui écrivirent sur la statistique à partir du XVIe siècle visaient avant tout la connaissance de l'Etat et ses ressources. Ce qu'ils créèrent, ce fut une science descriptive de l'Etat, dont tous les éléments étaient destinés à l'Etat lui-même. Ce n'était pas la population pour

elle-même que l'on étudiait, mais la population en tant qu'un facteur de la prospérité de l'Etat. Rentrent dans les travaux se rattachant à cette conception, la «Cosmographie» de Sébastien Muenster (1489-1552), le traité « del governo ed amministrazione » de Francesco Sansovino (1521-1586), les « relationes universales » de Giovanni Botero (1608), l'ouvrage de Pierre d'Avity (1572-1635), intitulé : « les Etats, Empires et Principautés du Monde », enfin, toute la collection des « Républiques Elzéviriennes » commencée en 1626, consacrée à la description des Etats. Cette soixantaine de petits volumes font aujourd'hui les délices des amateurs de livres anciens, mais, à les feuilleter, personne ne pourrait se douter qu'on se trouve en présence d'une œuvre qui passa autrefois pour appartenir à la statistique.

12. Un érudit, Herman Conring (1606-1681), médecin réputé, professeur à l'Université de Helmstädt, fut le premier à faire de ces descriptions politico-économiques un exposé systématique, sous la forme de cours, destiné à l'enseignement universitaire (1660). Il emprunte ses données aux publications antérieures. La matière qu'il embrasse dans ses descriptions est appelée par lui « Notitia rerum publicarum ». L'exposé qu'il fait est basé sur la méthode aristotélicienne; il commence par décrire le « comment » des choses : « quemadmodum res alicujus reipublicae sese habeant, ut Aristoteles solet loqui τοῦ ὅτι »; ensuite il examine les causes de l'état de choses constaté : « quibus de causis res istae hunc aut illum in modum se habeant, quod est το διότι ». Pour connaître pleinement la situation de chaque Etat, il faut, dit Conring, sur ce point fidèle disciple d'Aristote, connaître les quatre causes suivantes : 1° *causa materialis*. C'est ce qui se rapporte à l'esprit, au corps, à la fortune. Pour cette dernière, il faut distinguer les biens mobiliers et les biens immobiliers; parmi ces derniers se trouve la terre elle-même; 2° *causa finalis*, c'est-à-dire si les habitants peuvent vivre heureux, dans la pra-

tique des vertus et s'ils ont des moyens d'existence suffisants; 3° *causa formalis,* ou manière dont le pays est gouverné; 4° *causa efficiens,* autrement dit : « omnes qui regunt ». Cette *causa efficiens* est double; comme le fait remarquer Conring, l'une est principale, c'est l'homme qui gouverne; l'autre est instrumentaire et elle se divise en deux : *causa animata* (ministres, magistrats), et *causa non animata* (en premier lieu l'argent). C'est aussi à cet endroit qu'il convient de parler de l'armée et de la navigation.

Ces détails montrent combien l'enseignement se modèle encore sur les formules anciennes et font saisir, en même temps, combien peu cet enseignement se rapproche de ce qu'il est convenu aujourd'hui de désigner sous le nom de statistique. Conring n'a pas laissé d'exposé écrit de son œuvre, mais ses leçons furent recueillies et publiées par ses disciples. Le nouveau mode d'exposer cette « notitia rerum publicarum » fut suivi par de nombreux professeurs en Allemagne. Il devait appartenir à Achenwall de pousser à un degré supérieur cette systématisation de la description politico-économique des Etats.

13. Gottfried Achenwall (1719-1772), professeur à l'Université de Marburg, puis à celle de Göttingen, continua l'œuvre de Conring; mais, d'après Wagner, il donna à la statistique une entière indépendance, il en détermina avec plus d'exactitude les limites et l'objet, il la popularisa comme science en lui donnant un nom bien à elle, il lui assigna un domaine plus étendu et établit, mieux que tout autre avant lui, les relations étroites de la statistique avec la vie publique. Tandis que les travaux de Conring ne furent guère connus que dans des milieux scientifiques assez limités, ceux d'Achenwall conquirent la notoriété. Les érudits allemands se sont plu à lui décerner le titre de « père de la statistique ». Ceci est fort exagéré. Achenwall est le père du *nom,* mais pas de la science. S'il y avait quelque utilité à cette recherche des « priorités », on trou-

verait sans peine les origines de la science en Angleterre au XVII[e] siècle, au lieu de lés rechercher en Allemagne au XVIII[e].

C'est dans l'introduction de son œuvre principale, publiée en 1749, intitulée : « Abriss der Neusten Staatswissenschaft der heutigen vornehmsten europäischen Reiche und Republiken », qu'il employa pour la première fois le terme de « statistik », dérivé de l'italien « statista ». Achenwall faisait remarquer que parmi les faits qui se passent sous nos yeux, certains intéressent grandement la prospérité publique, en accélèrent ou en retardent la marche. Ce sont ces faits qui constituent vraiment les choses remarquables de l'Etat dont la connaissance constitue le domaine propre de la statistique. L'intérêt de l'Etat est interne et externe ; les moyens d'augmenter la richesse, de stimuler la population, de promouvoir les sciences, l'industrie et le commerce visent l'intérêt interne ; les alliances et les relations étrangères se rattachent à son intérêt externe.

La statistique chez Achenwall est donc une description de l'Etat, elle est une science purement descriptive. Il ne vient pas à la pensée d'Achenwall de déduire des faits observés des règles générales, des lois visant la marche des phénomènes qu'il observe. Son ambition se borne à fournir des éléments qui, par leur comparaison, permettront à l'homme d'Etat d'arriver à un plus haut degré de sagesse politique. Ses travaux de statistique descriptive ont porté sur l'Espagne, le Portugal, la France, l'Angleterre, les Pays-Bas, la Russie, le Danemark et la Suède. L'Allemagne, la Prusse et l'Autriche ne furent pas étudiées parce que, à l'époque d'Achenwall, la description de l'Etat était considérée, dans ces pays, comme une branche du Droit public, lequel avait son enseignement séparé. L'exposé d'Achenwall dérive des travaux publiés à son époque, il ne découle pas de recherches originales et ne fait pas la critique des documents qu'il analyse ; par sa clarté, sa concision, et le soin qu'il apporte dans son exposé, Achenwall est un modèle.

Son œuvre fut traduite dans toutes les langues, ce qui contribua à répandre le nom de « statistique » que le professeur de Göttingen avait donné à la discipline nouvelle.

14. Achenwall fit école et pendant de longues années ce fut sa conception de la statistique qui domina. De nombreux professeurs exposèrent la statistique d'après ses vues, de non moins nombreux manuels parurent qui suivaient fidèlement sa doctrine. Sans doute, on peut relever d'assez fréquentes divergences avec les vues du professeur de Göttingen, mais, dans l'ensemble, l'école d'Achenwall reste attachée à la notion de science descriptive des choses les plus remarquables de l'Etat. Mais à quel signe reconnaître ces « choses remarquables » ? Un des plus fidèles disciples d'Achenwall, son successeur dans la chaire universitaire de Göttingen, Louis von Schlözer (1735-1809), se posant cette question, faisait remarquer que ces « choses remarquables » étaient, somme toute, une notion variable selon le temps, le lieu, les circonstances. Ce qui ne mérite pas attention aujourd'hui peut acquérir demain une réelle importance; considérée sous l'aspect de l'intérêt général, telle chose peut être dénuée de toute importance, mais elle peut devenir digne d'attention si on prend égard à tel intérêt spécial, particulier. Il fallait donc essayer de trouver un critère scientifique d'après lequel on pût déterminer plus exactement le rôle et les limites de la statistique. A la fin du XVIIIe siècle, on le voit apparaître sous la notion de la « force » ou des « ressources » de l'Etat. Tout ce qui regarde la puissance de l'Etat, c'est-à-dire l'étendue de son territoire, la population, l'agriculture, le commerce, l'industrie, la marine marchande, les forces de terre et de mer, constitue le domaine propre de la statistique. Et L. von Schlözer, donnant corps à cette conception générale, précisait l'objet propre de la statistique, à l'aide des mots : « vires unitae agunt », c'est-à-dire : « vires », les forces naturelles ou acquises (les hommes, le territoire, les avan-

tages naturels), « unitae », la forme et les organes du gouvernement, « agunt », le gouvernement et l'administration.

15. Avant que l'école de Conring-Achenwall-Schlözer se fondât et se développât en Allemagne d'abord, dans la plus grande partie de l'Europe ensuite, un courant scientifique s'était établi en Angleterre, dont l'orientation spéciale devait exercer sur la statistique une influence prépondérante. Il s'agit de l'école des « arithméticiens politiques », dont Graunt (1620-1674) et Petty (1623-1687) furent les représentants les plus autorisés. Leur méthode est basée sur l'observation numérique, mais comme à leur époque les recherches statistiques dans le sens où nous entendons cette expression aujourd'hui, sont rares, coûteuses, souvent même impossibles, c'est par des calculs approximatifs, l'emploi des proportions, des procédés parfois déconcertants, qu'ils s'efforcent de suppléer à l'insuffisance des données numériques. On a beaucoup critiqué, parfois ridiculisé, leurs procédés, mais au point de vue qui nous occupe à présent, leur méthode numérique est extrêmement intéressante parce que, perfectionnée et surtout adaptée aux faits, c'est elle qui va se substituer plus tard à la science purement descriptive d'Achenwall.

Nous nous trouvons bien en présence de la doctrine qui doit devenir la statistique telle que nous la connaissons et la pratiquons. Certes, ses moyens d'action sont limités, son langage est encore un balbutiement, mais que la technique se constitue, que les facilités matérielles lui soient acquises et elle ne tardera pas à parler haut et clair.

L'essence de la statistique se trouve définie par Petty lui-même dans la préface de son « Political Arithmetic » (1691) : « La méthode que je suis, dit-il, n'est pas encore fort usitée ; au lieu de me servir seulement de vocables comparatifs ou superlatifs, j'ai voulu m'exprimer au moyen de nombres, de poids et de mesures, me servir seulement d'arguments matériels et considérer uniquement les causes qui

sont visiblement fondées sur la nature en laissant de côté celles qui dépendent des sentiments, des opinions, des appétits et des passions variables des hommes. » Avant Petty, John Graunt avait déjà fait (1662) de curieuses recherches, basées sur les registres mortuaires de la ville de Londres, recherches dont le résultat fut présenté à la « Royal Society » qui venait d'être fondée en cette ville. Il fut le premier à attirer l'attention sur la prédominance des naissances masculines sur les naissances féminines, qu'il estimait être dans le rapport de 14 à 13, soit 107 p. c., alors que la statistique moderne fixe ce rapport à 105 p. c. environ, montrant ainsi que les calculs de Graunt n'étaient pas si éloignés de la réalité. De même, il calcula le nombre de décès par période d'âges et montra qu'à l'aide du nombre de décès on peut connaître le nombre de vivants. On doit rattacher à la même école et à la même tendance la formation par l'astronome Halley (1656-1742) de la première table de mortalité, basée sur les tables de naissances et de décès dressées entre les années 1687-1691 par le prébendier Gaspar Neumann, pour la ville de Breslau. Les travaux postérieurs sur l'assurance-vie et les annuités dérivent de ces premiers essais; ils se développèrent surtout en Hollande et en Angleterre. C'est vers la même époque (1713) que Jacques Bernouilli publia son travail sur le calcul des probabilités.

16. La statistique basée sur les chiffres au lieu de consister en une description plus ou moins exacte des ressources de l'Etat, reçut surtout une impulsion décisive à la suite de la publication en 1741 de l'œuvre de Johann Peter Süssmilch (1707-1767), intitulée « L'ordre divin dans les variations du genre humain, preuve évidente de la divine Providence », etc. Cette œuvre est basée exclusivement sur les registres diocésains de la Prusse que l'auteur, membre du consistoire, avait pu consulter aisément. L'œuvre de Süssmilch, s'appuyant sur des matériaux inédits et d'une

grande richesse, a une valeur scientifique incontestable ; le rapport de 21 naissances masculines à 20 naissances féminines est démontré par l'auteur, en même temps que l'égalité du nombre des individus des deux sexes à l'époque du mariage. De ce fait, Süssmilch conclut à l'ordre divin de la monogamie. Toute son œuvre consiste à tirer parti des grands nombres, à en déduire les tendances générales qui expriment les *normalités* ou *lois,* à montrer les obstacles qui contrarient le jeu normal de ces lois voulues par la Divinité et à en réclamer la disparition. Il a déjà une notion claire de l'analyse statistique et lorsqu'il traite une question, que ce soit celle de la mortalité ou celle de la nuptialité, il a toujours soin d'en examiner les différents aspects sous le triple rapport du temps, de l'espace et des qualités.

« Lorsqu'on voit, dit Süssmilch, que sur un nombre donné d'hommes, il y en a autant qui meurent une année que l'autre ; qu'en outre, en un temps donné, il meurt tant d'enfants, de jeunes gens, d'adultes, de vieillards ; que les deux sexes se reproduisent dans une proportion constante, on doit admettre qu'en toutes ces choses il existe un ordre, et même un ordre extraordinairement majestueux, beau et parfait. » Trouver l'application de cet ordre divin dans les phénomènes humains n'était pas difficile, mais il fallait que, comme Colomb a découvert l'Amérique, quelqu'un indiquât la voie à suivre : ce Colomb de la statistique, écrit Süssmilch, a été Graunt, qui, par ses observations dans les registres des morts et des malades, à Londres, avait été amené à conclure à l'existence de cet ordre, et à sa généralisation dans les autres phénomènes de la vie humaine.

L'œuvre de Süssmilch a une très grande importance.

Si l'on peut considérer l'œuvre d'Achenwall comme la personnification de la statistique descriptive, on doit voir en Süssmilch un des promoteurs de la statistique mathématique. La statistique chez Achenwall est une science descriptive de l'Etat, science qui vise spécialement les conditions sociales à un moment déterminé. Cette description

a une utilité pratique, tant pour l'administration que pour la politique. Süssmilch a une autre théorie. Pour lui, le matériel statistique et toutes les recherches statistiques n'ont pas seulement pour but de faire connaître l'Etat et les conditions dans lesquelles il se trouve. Il veut que l'on recueille des données systématiquement présentées, à l'aide desquelles on puisse trouver l'explication de certains phénomènes de la vie humaine et indiquer les lois qui régissent ces phénomènes. Il est incontestable que cette conception est plus élevée et plus féconde en résultats scientifiques que celle d'Achenwall.

17. Nous avons vu s'établir un double courant dans la conception de la statistique, dans le but qu'elle poursuit, dans les matériaux qu'elle met en œuvre, dans la forme sous laquelle elle présente ses résultats. L'école de Conring-Achenwall-Schlözer, l'école de Göttingen, d'une part, préconise une science descriptive, appliquée aux choses remarquables de l'Etat, visant les phénomènes contemporains. L'autre école, celle des arithméticiens politiques, étend cette notion et cet objet, a recours aux données numériques et au calcul, recherche les rapprochements numériques, en tire des déductions, que ses adversaires estiment, pour le moins, hasardeuses. Les deux écoles se développèrent simultanément, parallèlement, mais leur antagonisme ne devait pas tarder à se révéler.

Ce serait cependant exagérer que de dire que l'école de Göttingen repoussait systématiquement et d'une façon absolue l'emploi des données numériques et les comparaisons. Sous ce dernier point de vue, il convient de mentionner ici le procédé comparatif auquel recourut le premier, en 1758, Antoine-Frédéric Büssching, géographe réputé, qui entreprit d'exposer la statistique sous forme de rubriques distinctes, dans lesquelles les éléments de même nature se trouvaient groupés pour plusieurs Etats. C'était un perfectionnement apporté à la méthode d'Achenwall,

et von Schlözer lui-même, qui s'était constitué le gardien intraitable des traditions de son maître Achenwall, ne put s'empêcher de proclamer excellente cette innovation. La méthode nouvelle recruta de nombreux imitateurs au début du XIXe siècle, mais la méthode de Büssching est, au fond, la même que celle d'Achenwall. Elle se limite toujours aux considérations sur la constitution, la force, les conditions matérielles de l'Etat et se sert, comme le fit l'école de Göttingen, de la méthode d'exposition de la science historique.

L'école d'Achenwall n'avait pas rejeté *ex professo* tout emploi des données numériques. Ses tenants se rendaient compte que, sur bien des points, les chiffres parlaient un langage d'une autre précision que les qualificatifs les mieux choisis. Suffit-il de noter qu'un pays produit de bons vins, possède de belles fabriques, a un commerce florissant pour donner au lecteur une impression exacte de ce que sont ces produits des vignobles, ces manufactures, ce commerce? Von Schlözer reconnaissait que des nombres feraient ici bien mieux l'affaire. Il écrivait lui-même : « Une donnée peut être exacte et cependant être, somme toute, inutile si on ne peut l'exprimer en chiffres ». Mais il ne s'agissait pas pour l'école de Göttingen de réduire toute la statistique à des chiffres ; tout au plus était-elle disposée à admettre que, pour un pays, on dressât un tableau numérique où se trouveraient quelques résultats importants.

Cependant, l'école de l'arithmétique politique avait continué à se développer et avait trouvé, en Allemagne même, d'assez nombreux adhérents et propagandistes. Dès l'année 1741, un Danois, P. Anchersen, avait même entrepris de comparer, au moyen de tableaux numériques, les conditions respectives de plusieurs Etats, procédé que l'école de Göttingen ne devait imiter, à l'aide de ses descriptions, que quelques années plus tard, en 1758, lors de la publication du livre de Büssching.

Le heurt était inévitable entre les deux écoles. Les fidèles

de la méthode d'Achenwall surtout y mirent de la passion et de l'ironie. Ils reprochèrent à leurs rivaux, bientôt devenus leurs adversaires irréductibles, de ravaler la statistique à un amas confus de chiffres sans sécurité et sans portée. Ils les appelaient les « fabricateurs de chiffres, les esclaves des chiffres ». Au commencement du XIXe siècle, la querelle avait atteint des proportions épiques en Allemagne. « La statistique, écrivaient des polémistes qui prenaient le parti de l'école de Göttingen, est devenue un travail insensé uniquement par la faute des arithméticiens politiques. Ces auteurs croient et s'imaginent faire croire que les ressources d'un Etat peuvent être connues en sachant uniquement quel est le nombre de milles carrés que représente le territoire du pays, le nombre de ses habitants, sa population relative, le revenu national, etc. »

Aussi l'œuvre des arithméticiens politiques leur paraissait-elle néfaste : elle tendait, disaient-ils, à réduire la statistique à l'état de squelette, à la priver de vie et à la dépouiller de sa physionomie propre. Au contraire, l'école de Göttingen prétendait conserver à la statistique les cadres si vastes tracés par Achenwall et von Schlözer, qui s'étendaient à bien des matières que les chiffres ne pouvaient exprimer et, notamment, à des questions psychologiques et politiques que, seule, la description littéraire pouvait aborder. Il y avait une part de vérité dans ces critiques, à côté d'énormes exagérations. Il était exact que les données numériques auxquelles les arithméticiens politiques avaient recours ne présentaient pas toutes, il s'en fallait de beaucoup, le caractère d'exactitude qu'il eût été désirable de leur voir revêtir ; il est vrai aussi que ces données étaient en trop petit nombre pour prétendre exprimer l'aspect social et économique d'un pays. Mais, par contre, il était injuste de décrier des efforts qui allaient ouvrir à la statistique une voie féconde. Et de plus, le point de vue auquel se plaçait l'école de Göttingen, défendant jalousement la situation qu'elle avait acquise dans l'enseignement acadé-

mique, était faux. Il procédait d'une conception erronée du but de la statistique et d'une méconnaissance des moyens d'y parvenir. L'école d'Achenwall faisait un étrange amalgame de la géographie, de l'histoire, de l'économie politique, voire de la morale et de la psychologie, avec la statistique proprement dite. L'esprit scientifique, en se développant, devait aboutir, comme il aboutit en fait, à dissocier ces éléments disparates. L'économie politique se constitua en science autonome avec l'œuvre d'Adam Smith : « La richesse des nations » (1776) ; la philosophie développa la notion de la nature et de la fonction de l'Etat ; les sciences administratives occupèrent une place à part dans l'enseignement universitaire ; la géographie se plaça sur son véritable terrain en étudiant spécialement les conditions naturelles résultant du climat, de l'orographie, de la géologie, et de l'ethnographie. La science de l'actuariat adopta comme base les mathématiques et ne fit plus appel à la statistique que pour lui réclamer des éléments de travail. Bref, l'ancienne statistique descriptive se vit privée de la plupart des domaines sur lesquels elle s'étendait autrefois, tandis que la statistique numérique n'était pas encore arrivée à un état de développement scientifique qui permît d'en faire l'objet d'un enseignement académique. Il semblait donc que l'arrêt du développement scientifique de la statistique était complet et définitif et que la science, à peine née, allait disparaître.

Mais on peut appliquer à un nouveau venu l'hémistiche que Boileau écrivait à propos de Malherbe, et dire : « Enfin, Quetelet vint... ».

18. Lambert-Adolphe-Jacques Quetelet, né à Gand le 22 février 1796, décédé à Bruxelles, le 17 février 1874, directeur de l'Observatoire royal de Bruxelles, secrétaire perpétuel de l'Académie royale de Belgique, président de la Commission centrale de statistique de Belgique, eut une vie d'une telle activité scientifique qu'il faudrait tout un

volume pour la retracer (1). Nous avons déjà dit notre intention d'indiquer quelques sommets seulement parmi la longue chaîne d'auteurs qui se sont appliqués à cette discipline, sans élever aucune prétention à faire l'histoire de la statistique, histoire à laquelle ont été consacrés des travaux spéciaux d'une grande valeur. Il s'agit uniquement de spécifier comment un certain nombre d'auteurs entendirent la statistique et de montrer en quoi ils en augmentèrent la connaissance. C'est à ce point de vue unique que nous nous placerons dans ce court résumé de l'œuvre de Quetelet, laissant de côté ses théories sur la mécanique sociale et sa conception de l'homme moyen qui exigeraient de trop longs développements.

Dans ses premiers travaux, Quetelet ne semble pas se former de la statistique une opinion très différente de celle qui était courante à son époque. Dans le second travail statistique qu'il fait paraître, les « Recherches statistiques sur le Royaume des Pays-Bas », ouvrage qui fut publié en 1828, Quetelet écrit que la statistique est l'anatomie de la société, que son but est « d'observer les modifications qu'éprouvent les différents peuples dans leur état physique et moral et de chercher d'en pénétrer les motifs » (2). Il convient de remarquer cependant que cette définition renferme l'idée de comparabilité et celle de la recherche des causes. Or, si la première a été appliquée par Büssching, la seconde, celle de la recherche des causes, a toujours été étrangère à l'école d'Achenwall, qui réservait aux sciences historiques l'explication des faits qu'elle enregistrait. Sans aller jusqu'à dire que Quetelet innove et est en avance sur

---

(1) Cf. Frank H. Hankins, *Adolphe Quetelet as statistician* (Studies in history, economics and public law edited by the Faculty of political science of Columbia University. New-York, Columbia University, 1908). Joseph Lottin: *Quetelet statisticien et sociologue* (Bibliothèque de l'Institut supérieur de philosophie de l'Université catholique de Louvain). Louvain-Paris, 1912. Cette étude contient la bibliographie de Quetelet et la liste des principaux ouvrages sur Quetelet.

(2) Quetelet, op. cit., p. 11, « Introduction ». Bruxelles, Hayez, 1846.

son temps — car le programme de l'école de Göttingen avait, au cours des années, subi de nombreuses modifications — il est bon de signaler que notre auteur, dès le début de sa carrière, est à l'avant-garde.

La pensée de Quetelet n'est pas toujours facile à démêler parce qu'il lui donne souvent une expression différente et qui, pour le fond, ne semble pas toujours adéquatement la même. Il faut aussi, parmi ses nombreux écrits, choisir celui qui paraît le plus convenable pour y rechercher les idées de l'auteur. Or, au point de vue où nous envisageons l'œuvre de Quetelet, cet ouvrage nous paraît être les « Lettres sur la théorie des probabilités appliquées aux sciences morales et politiques » ; d'abord parce que Quetelet a écrit cette œuvre alors qu'il était dans toute la force de son talent : cet ouvrage a été commencé en 1837, l'auteur avait alors 41 ans ; ensuite, parce que les « Lettres sur la théorie des probabilités » renferment une série systématique de considérations sur la statistique. Quetelet n'avait pas, disait-il, la prétention de donner un traité sur cette matière, il s'est simplement borné à signaler plusieurs erreurs dans lesquelles les auteurs tombent fréquemment, et à jeter en avant quelques idées qui lui ont paru propres à éclaircir la théorie (1). Cela n'empêche que l'exposé fait par Quetelet présente l'allure systématique d'un exposé doctrinal où il est permis de chercher l'expression des idées de l'auteur.

19. La statistique, d'après Quetelet, a pour sujet l'Etat, c'est-à-dire une agrégation d'hommes possédant un territoire et un gouvernement (2).

Ce n'est pas uniquement la société politique qui est à considérer, mais la société civile, car l'homme, être sociable, est l'élément essentiel de l'Etat (3).

---

(1) QUETELET, *loc. cit.*, « Introduction », p. IV.
(2) *Ibid.*, p. 260.
(3) *Ibid.*, p. 260.

Faire de la statistique, c'est considérer un Etat pendant une de ses phases de développement, à un moment donné, pour reconnaître son organisation avec tout ce qui l'entoure (1).

La définition de l'objet de la statistique est la suivante: « Elle ne s'occupe d'un Etat que pour une époque déterminée; elle ne réunit que les éléments qui se rattachent à la vie de cet Etat, s'applique à les rendre comparables et les combine de la façon la plus avantageuse pour reconnaître tous les faits qu'ils peuvent nous révéler » (2).

Mais quels sont les éléments qui se rattachent à la vie de l'Etat ? L'école de Göttingen n'avait, par exemple, attaché qu'une importance secondaire aux éléments économiques, mais une grande aux éléments politiques.

Quetelet envisage la vie de l'Etat dans le sens le plus étendu. D'après lui, la statistique générale d'un Etat comprend essentiellement les cinq divisions suivantes: 1° la population; 2° le territoire; 3° l'état politique; 4° l'état agricole, industriel et commercial; 5° l'état intellectuel, moral et religieux. La population, dit Quetelet, est l'élément statistique par excellence: il amène nécessairement tous les autres, puisqu'il s'agit, avant tout, du peuple et de l'appréciation de son bien-être et de ses besoins (3).

L'homme est le sujet vrai de la statistique. Ainsi, en traitant du territoire, ou de l'agriculture, ce n'est pas la composition du sol qu'il faut avoir en vue, mais seulement les objets qui servent aux usages de l'homme, soit pour être consommés immédiatement, soit pour être utilisés par le commerce et l'industrie (4).

Si l'on s'occupe des animaux, ce n'est que pour autant qu'ils sont utiles ou nuisibles à l'homme (5).

---

(1) QUETELET, *loc. cit.*, p. 260.
(2) *Ibid.*, pp. 268-69.
(3) *Ibid.*, p. 270.
(4) *Ibid.*, p. 274.
(5) *Ibid.*, p. 275.

Or, ce point de vue est essentiel, car il permet d'éviter l'écueil qu'ont rencontré certains statisticiens qui ont transporté dans la statistique d'autres sciences qui lui sont étrangères, comme la géographie physique, la minéralogie, la botanique, la météorologie, etc. (1)

Ces éléments, comment peut-on les représenter par des chiffres ou par une description littéraire ? C'était là, on se le rappelle, une distinction fondamentale entre la doctrine des arithméticiens politiques et celle de l'école de Göttingen. Quetelet se montre éclectique, et ici, il prévoit clairement les nécessités de la statistique dans son développement ultérieur. Certains auteurs, dit-il, au lieu d'étudier la statistique, sont tombés dans le travers opposé et ont voulu la restreindre, la réduire à ne présenter que des tableaux purement numériques, sans songer qu'il existe des renseignements qu'il serait impossible d'exprimer en nombres (2). L'exposé de l'état politique, par exemple, appartient essentiellement à la statistique d'un pays, et cependant on ne saurait le faire connaître par des chiffres. On peut en dire autant de beaucoup de renseignements relatifs à l'état moral et intellectuel. Le simple récit de ce qui s'est passé dans une localité à une époque donnée en apprend quelquefois plus sur l'état moral d'un peuple que tous les tableaux numériques possibles (3).

La statistique ne se confond-elle pas avec l'histoire et la politique ? Non, répond Quetelet. La statistique est à l'histoire « ce que, dans un ordre de choses différent, la statique est à la dynamique, ce que le repos est au mouvement » (4). Comment ne pas se souvenir de la définition de von Schlözer : « L'histoire est une statistique en mouvement, la statistique est une histoire en repos ». Quetelet

(1) Quetelet, *loc. cit.*, p. 275.

(2) *Ibid.*, p. 275.

(3) *Ibid.*, p. 276.

(4) *Ibid.*, p. 261.

ajoute : « plus généralement, la statistique s'occupe de l'instant présent, en laissant le passé à l'histoire, et l'avenir à la politique » (1).

Le sujet de la statistique, a dit Quetelet, c'est l'Etat, c'est-à-dire les hommes qui composent la société civile. Tel est le domaine propre de la statistique. Mais, en même temps, Quetelet fait des applications constantes des procédés statistiques à la météorologie, à l'astronomie, à la botanique, mais il distingue soigneusement de la statistique toutes les sciences qui lui sont étrangères. C'est donc qu'il admet, encore qu'il ne l'ait pas expressément formulée, la distinction entre la science statistique et la méthode statistique.

20. L'influence de Quetelet se manifeste dans quatre domaines importants, à propos : 1° de la statistique de la population ; 2° de la statistique morale ; 3° du développement de la technique ; 4° de l'application de la loi normale des erreurs aux facultés de l'homme. Peut-être n'a-t-il pas innové complètement en aucune de ces matières, mais en toutes il s'est montré original, car sa conception est plus haute, plus claire, plus scientifique que celle d'aucun de ses devanciers. Quetelet a rendu un service inappréciable à la science en combinant harmonieusement les vues des arithméticiens politiques et celles de l'école d'Achenwall. Il est arrivé à une époque où les recherches numériques se sont suffisamment développées pour fournir des éléments nombreux et divers. Il n'a cessé de promouvoir ces recherches, d'en perfectionner la technique, de recommander l'emploi d'une critique judicieuse et attentive. Dans toutes ses investigations, il a toujours été aidé d'un grand bon sens. S'il s'est servi des chiffres, il n'a pas été leur esclave et a eu recours à toutes les ressources de l'analyse logique. Bref, il a donné un corps à une foule d'éléments épars, il a per-

(1) QUETELET, *loc. cit.*, p. 261.

fectionné toutes les matières auxquelles il a touché: il a constitué la statistique moderne.

II. — Diverses opinions en présence.

21. D'après les diverses conceptions modernes, à quel sujet s'attache la statistique, qu'est-elle comme objet de connaissance scientifique, quelles sont ses limites, quel est son contenu logique ? Ce sont les questions que Meitzen pose comme entrée en matière à l'un de ses chapitres les plus remarquables et qu'à notre tour nous nous posons après la revue sommaire qui vient d'être faite.

Un premier groupe se présente à nous : celui d'Achenwall et de son école. Nous faisons abstraction pour le moment de son éloignement pour la donnée-chiffre, qui, du reste, diminua avec le temps, pour ne considérer que le sujet attribué par cette école à la statistique. Ce sujet est l'Etat et ses propriétés remarquables. « La statistique est la connaissance approfondie de la situation respective et comparative de chaque Etat », ou encore: « l'ensemble de ce qui est réellement remarquable dans un Etat et en forme la constitution dans le sens général forme le domaine de la statistique ». L'important aux yeux d'Achenwall et de ses successeurs, c'est de préciser le sujet des observations de la statistique. Les questions de méthode, pour eux, sont accessoires. L'influence de l'école de Göttingen — nous l'avons dit plus haut — alla en diminuant, mais au cours de la première partie du XIX[e] siècle, on trouve encore des auteurs de talent qui font abstraction du contenu de la statistique et de son essence, pour considérer uniquement le sujet de ses études. « La statistique, écrit Wappäus en 1859, doit encore aujourd'hui s'appuyer sur le concept d'Achenwall si elle ne veut pas perdre tout à fait le caractère de science positive. » D'après cet auteur, la statistique n'a pas un caractère phi-

losophique, c'est une science positive qui comprend un complexus de connaissances subordonné à un but pratique précis. Ce but est la connaissance d'un Etat déterminé, concret, tel qu'il est dans la réalité. La statistique comprend la statistique comparative, laquelle conduit à une statistique scientifique quand elle est exprimée en chiffres pouvant se transformer en rapports, etc. La statistique de l'Etat serait la statistique *spéciale*. L'autre serait la statistique *générale* comparée, ayant un caractère philosophico-mathématique.

Dans cette conception, ce qui importe le plus à la statistique, c'est l'Etat. L'homme n'est pas au premier plan, il ne vient comme sujet d'étude que parce qu'il est partie intégrante et agissante dans l'Etat. La statistique alors se range parmi les sciences politiques et administratives.

22. Mais la statistique ne devait pas tarder à élargir ses frontières traditionnelles. Lorsque Quetelet, dans ses premiers travaux, faisait servir la statistique à déterminer ce qu'il y a de normal ou de plus fréquent parmi les caractères physiques et moraux de l'homme, il mettait en avant un principe fécond dont l'application méthodique devait orienter la statistique vers des voies nouvelles. Les mémoires sur les lois des naissances et de la mortalité à Bruxelles (1825), les recherches statistiques sur le Royaume des Pays-Bas (1828), son mémoire sur la constance qu'on observe dans le nombre des crimes qui se commettent (1830), les recherches sur la loi de croissance de l'homme (1831), les études sur le penchant au crime aux différents âges (1831), celles sur le poids de l'homme (1832), préparaient les deux volumes que Quetelet publia en 1835 avec le titre : « Sur l'homme et le développement de ses facultés ou Essai de physique sociale ». En montrant que la loi des erreurs se vérifiait dans les qualités physiques et morales de l'homme, Quetelet ouvrait un vaste champ aux recherches des statisticiens fatigués de se cantonner dans les descrip-

tions de l'école de Göttingen, dans un champ de recherches jalousement borné par la politique, l'histoire, la géographie et l'éthique. En même temps, ces travaux de Quetelet mettaient un terme à la vieille controverse entre les arithméticiens politiques et les disciples d'Achenwall, car si les recherches nouvelles inaugurées par lui étaient impossibles sans de nombreuses données numériques, elles ne pouvaient non plus avoir de signification sans les commentaires qui les accompagnaient. Ainsi se fusionnaient, dans une conception véritablement scientifique, les prétentions opposées des deux écoles. Du coup, le domaine de la statistique recevait une extension pour ainsi dire illimitée, en même temps que l'enseignement universitaire, si développé en Allemagne, prenait une direction nouvelle.

L'orientation donnée à la statistique prit le dessus rapidement; au milieu du XIX[e] siècle, écrit Meitzen, elle était devenue générale.

23. Parmi les auteurs qui développaient ce point de vue, nous en choisissons quelques-uns, à titre d'exemple, pour montrer comment la statistique est comprise par ceux qui adoptèrent les idées émises par Quetelet quant à l'objet de la statistique. Lexis considère comme appartenant au domaine de la statistique envisagée comme science, toutes les recherches qui se rapportent aux faits de la vie humaine qui peuvent être observés au moyen du relevé en masse. « La statistique, dit-il, a pour office propre de recueillir et d'étudier, d'après une méthode exacte, les faits de la vie humaine, pris en masse et de cette définition il suit déjà que la base de la méthode consiste dans le comptage des cas individuels d'un phénomène. Elle ne considère pas, comme l'histoire, l'individualité des événements, mais elle les enregistre seulement comme des éléments d'une masse, comme des unités d'un total. »

Del Vecchio exprime la même idée sous une forme concise: « La statistique moderne est une science positive. dé-

terminée par son objet, qui est la société humaine et par sa méthode, qui consiste dans l'observation d'une quantité de phénomènes homogènes et sur une grande échelle ou plus brièvement, comme disent d'autres définitions, dans l'observation en masse ».

Pour von Mayr, le domaine de la statistique scientifique comprend l'étude : 1° des masses humaines ; 2° des masses de faits humains et d'événements sociaux ; 3° des masses d'effets d'actes humains et d'événements.

L'exposé de Rumelin est surtout remarquable. Il importe cependant de faire remarquer que dans la distinction entre les phénomènes sociaux et naturels, le statisticien allemand ne faisait que reprendre les idées émises par Dufau (1) vingt-trois ans auparavant.

Rumelin part de ce principe que « le particulier est typique dans le règne naturel, individuel dans le règne humain ». Comment des sciences, qui ont pour objet le règne humain, pourront-elles employer la voie de l'expérience ? Tant qu'elles restent sur le terrain des observations isolées, soit du présent, soit du passé, elles ne peuvent guère dépasser la sagesse des proverbes. Pour leur tenir lieu de ces instruments et de cette épreuve que possèdent les sciences du règne de la nature, il faut que les sciences du règne humain en arrivent à l'observation méthodique des masses. Celle-ci consiste à étendre sur des groupes entiers d'individus comme un réseau d'observations qui contemplent et enregistrent, d'après une méthode unique, tous les phénomènes semblables. De là une science auxiliaire s'est formée qui a pour tâche « la découverte des caractères des communautés humaines par l'observation méthodique et le calcul des phénomènes semblables » ; et par communautés, Rumelin entend non seulement les groupes naturels d'individus comme les peuples, les Etats, les provinces, etc., mais aussi

(1) DUFAU, *Traité de statistique*. Paris, 1840.

tous les autres cercles d'activité susceptibles d'un examen particulier, comme les relations politiques, économiques, sociales, religieuses, etc. (1). Cette science auxiliaire de toutes les sciences expérimentales du règne humain, c'est la statistique.

Ces extraits pourraient être multipliés sans autre résultat que de lasser le lecteur et d'allonger sans utilité cet exposé. Tels quels, ils suffisent à montrer qu'au cours du XIXe siècle la statistique a conquis un ordre de choses nouveau et que son domaine propre, ce n'est plus l'Etat et ses propriétés remarquables, mais l'homme lui-même, ce qui lui assure un champ d'action autrement vaste et intéressant.

24. Ce que nous venons de dire de Rumelin nous fournit naturellement l'occasion de passer à un nouvel exposé : celui des vues de l'école qui considère la statistique comme une méthode ou plus exactement comme une des sciences de la méthode. Rumelin, on l'a vu plus haut, base son argumentation sur ce point essentiel : le cosmos ou le monde se présente à nous sous deux grandes faces : le règne de la nature et le règne humain. Les sciences de la nature comme les sciences de l'homme sont des sciences expérimentales, en ce qu'elles reposent en dernière analyse sur l'induction et l'observation, peu importe d'ailleurs que le procédé déductif y joue aussi plus ou moins son rôle. Mais ces deux grandes classes de sciences n'en sont pas moins fort différentes quant à leurs moyens d'observation scientifique.

Les progrès des sciences naturelles s'expliquent par ce fait que, dans la nature, tout cas particulier peut servir de type, en sorte qu'un seul fait bien observé autorise déjà une

---

(1) RUMELIN, *Problèmes d'économie politique et de statistique.* Trad. franç. de A. de Riedmatten. Paris, Guillaumin, 1896, pp. 88-95.

induction, et qu'on n'y répète d'ordinaire l'observation que pour mieux en contrôler l'exactitude (1).

Cette thèse de Rumelin a rencontré, sinon des contradictions absolues — ce qui paraîtrait difficile — au moins des réserves. Wagner, dans son grand article *Statistik,* a fait remarquer qu'entre la nature et l'homme moral il n'y a qu'une différence de degré, non une différence principielle dans le système des causes. Ce système est plus compliqué chez l'homme, et dès lors il sera plus difficile de le démêler ; mais il ressemble au système des causes constantes et accidentelles, tel qu'il agit dans les phénomènes non typiques de la nature.

D'où cette conclusion que l'emploi de la méthode d'observation de la masse sera moins aisé dans le premier cas, mais qu'il est nécessaire dans les deux (2).

C'est, par réaction, aller trop loin. Rumelin, dans sa première étude, publiée en 1863, avait déjà pris soin d'apporter un correctif à ce que son affirmation pouvait paraître avoir de trop absolu. « Notre assurance, dit-il, diminue déjà quand nous passons aux plantes et aux animaux qui subissent l'action de l'homme, et elle s'évanouit tout à fait quand nous entrons dans le domaine de l'âme humaine (3). »

Dans une seconde étude, publiée en 1874, Rumelin précise ses vues à cet égard et introduit une distinction importante entre la statistique et la méthode statistique. « La statistique, dit-il, est une branche des sciences sociales et on doit la considérer comme telle, peu importe la définition qu'on en donne. Par contre, cette méthode particulière de recherche, dont le trait essentiel est de compter et de classer la multitude des cas, fut bien d'abord mise au service de la statistique et, par suite, des intérêts de l'Etat ; mais elle

(1) Rumelin, *loc. cit.*, pp. 86-87.

(2) Lottin, *loc. cit.*, p 256.

(3) Rumelin, *loc. cit.*, p. 87.

n'en est pas moins dans son essence et ses applications d'une portée beaucoup plus générale nullement limitée à ce cercle d'observations. Elle a, comme toutes les formes de méthode scientifique, sa place dans la logique. » (1)

L'ancienne logique ne connaissait que l'induction par laquelle elle recherchait les qualités qui appartiennent constamment à tous les individus, mais elle ne savait comment apprécier scientifiquement les éléments variables. Or, la méthode statistique a pour attribut essentiel de pouvoir, par l'observation scientifique et le calcul, caractériser les éléments variables des objets observés, pour les élever au rang de caractères utiles à la science. Elle révèle que des lois régulières gouvernent aussi ce domaine des variables, qu'il n'y a pas ici hasard ou arbitraire, mais seulement une plus grande complexité de forces et de causes (2).

« Et cette méthode, continue Rumelin, peut intervenir partout où les phénomènes observés présentent des variables, ce qui se rencontre dans tous les règnes de la nature. Partout aussi un intérêt scientifique peut se rattacher à cette variabilité. On pourrait compter et assortir utilement suivant leur grosseur et leur forme, jusqu'à des grains de sable, et l'on ne saurait encore prévoir toute l'importance qu'une supputation exacte des variables peut acquérir dans les sciences naturelles (3). »

25. De ce que la statistique apparaît, sous certain aspect, comme une méthode, faudrait-il conclure que tout ce à quoi elle s'applique comme méthode rentre dans le domaine d'une science unique, à laquelle on donnerait le nom de statistique? Ce serait soutenir une opinion fort hasardée et qui, par ses conséquences, confine à l'absurde.

---

(1) Rumelin, *loc. cit.*, p. 135.
(2) *Ibid.*, p. 136.
(3) *Ibid.*, p. 137.

Portlock, en 1838, soutenait cette opinion : « On peut dire, écrit-il, que le « statiste » est celui qui recueille des faits, que la statistique consiste dans les faits ou les données relatifs à toute chose ou à toute science, naturelle ou politique; que la science statistique est la réunion et l'ordonnancement de ces faits (1). »

Cournot, dominé par la préoccupation des caractères mathématiques de la statistique, ne pense guère autrement : « On entend principalement par statistique (comme l'indique l'étymologie) le recueil des faits auxquels donne lieu l'agglomération des hommes en sociétés politiques; mais pour nous, le mot prendra une acception plus étendue. Nous entendrons par statistique la science qui a pour objet de recueillir et de coordonner des faits nombreux dans chaque espèce, de manière à obtenir des rapports numériques sensiblement indépendants des anomalies du hasard, et qui dénotent l'existence des causes régulières dont l'action s'est combinée avec celle des causes fortuites (2). »

Nous n'entrerons pas dans le détail de l'examen de cette théorie, n'ayant pas ici à faire l'histoire des doctrines. Notons simplement que cette méconnaissance des caractères de la méthode et de la science contribua malheureusement à répandre au sujet de la statistique et de son objet les idées les plus fausses. C'est à elle également qu'il faut imputer les innombrables brocards dont la statistique fut l'objet durant de longues années.

Des vues plus sages se répandirent par la suite, notamment celle qui consiste à distinguer dans la statistique deux aspects : une science statistique qui s'applique aux phénomènes sociaux, une méthode statistique ayant pour objet

---

(1) *An address explanatory of the objects and advantages of statistical enquiries*. Belfast, 1838.

(2) COURNOT, *Exposition de la théorie des chances et des probabilités*. Paris, 1843, p. 181.

l'étude des phénomènes variables, applicable aux sciences naturelles comme aux sciences sociales.

Cette pensée a été exprimée clairement par von Mayr, qui propose une double définition : celle de la méthode statistique, ou statistique au sens formel, qui est l'observation quantitative numérique complète des phénomènes sociaux et de quelques phénomènes d'un autre ordre; et ensuite celle de la science de la statistique, ou statistique au sens matériel, qui consiste en l'explication des conditions et des phénomènes de la vie sociale, qui, comme tels, se manifestent dans les masses sociales, explication basée sur une observation numérique en masse, exprimée en nombres et mesures.

C'est une opinion encore fréquemment admise aujourd'hui.

Elle donne une extension nouvelle à la notion de statistique : celle-ci devient double et s'étend non seulement aux communautés humaines, mais, de plus, comme méthode, aux phénomènes de l'ordre des sciences naturelles.

26. Cette conception double — méthode et science — préparait elle-même une évolution de la pensée scientifique quant à la nature de la statistique. C'est qu'on pouvait lui reprocher divers inconvénients. Le premier était de scinder l'unité de la science : la statistique apparaissait tantôt comme une suite de raisonnements tendant à l'explication de phénomènes d'une importance capitale, c'est-à-dire comme une science ayant son objet et son but propres, tantôt comme une série de procédés utiles, sans doute, à la connaissance scientifique, mais dont le rôle, surtout dans les sciences naturelles, restait accessoire et inférieur, somme toute, à l'induction.

Etait-il raisonnable, d'autre part, de désigner sous le même nom deux disciplines ayant un objet distinct et un but différent : la statistique — science expliquant les phé-

nomènes, — la statistique — méthode accumulant seulement des matériaux?

Ou, s'il s'agissait d'une discipline unique, était-elle philosophique, cette conception d'une discipline à deux faces : science quand elle envisageait des phénomènes d'un ordre donné, méthode quand elle considérait d'autres aspects de l'univers?

Les logiciens et les mathématiciens se sont surtout attachés à considérer la statistique méthodologique et ils en ont fait une branche de la logique appliquée.

Un logicien allemand, Sigwart, a, le premier, développé ce point de vue d'une manière complète. Les jugements, dit-il, sont basés sur des perceptions par lesquelles les phénomènes sont décrits complètement et avec précision et situés dans le temps et l'espace. Là où cette description n'est pas possible, on y supplée par l'énumération statistique des phénomènes semblables en se basant sur la classification existante des objets considérés. La description des phénomènes individuels sert à les inclure dans les catégories déjà formées, mais par là on renonce à la description de l'individuel comme tel, car il est confondu dans un groupe d'unités semblables. Le caractère essentiel des relevés statistiques consiste en ce que les objets particuliers ne sont pas énumérés et catalogués comme tels, mais qu'ils constituent ensemble des totaux d'objets et de phénomènes semblables, en confondant les perceptions individuelles sous des rubriques distinctes. Ceci constitue la statistique descriptive qu'on peut considérer comme la première phase de la statistique même. La seconde phase consiste dans l'usage des résultats ainsi obtenus, dans le but de découvrir une loi.

Pour arriver à un pareil résultat, on a recours au calcul sous ses différentes formes. Les moyennes montrent l'existence de régularités empiriques; elles sont descriptives de leur nature et sont incapables d'exprimer une con-

trainte sans cette supposition que les circonstances qui produisent les cas variables sont constantes dans leur ensemble. Les conclusions statistiques sur la causalité sont fondées sur les variations des nombres et non sur leur constance. La statistique montre que les causes, connues par ailleurs, ont exercé leurs effets, qu'elles n'ont pas été annihilées par d'autres, et elle donne une mesure de l'intensité de chaque force agissante, en rapport de toutes les autres. Mais il est impossible à la statistique de prévoir et de dire que chaque cas individuel qu'elle englobe dans ses énumérations est soumis à la contrainte de la loi. Lorsque la loi est trouvée, l'analyse et la méthode inductive reprennent leur empire et la méthode statistique perd son intérêt.

Le statisticien russe Tchouprow base également le caractère logique de la statistique sur la nécessité de la connaissance de l'individuel et sur l'impossibilité d'arriver à cette connaissance autrement que par le moyen de la méthode statistique. Il distingue les sciences nomographiques qui fournissent les comptes généraux, des sciences idéographiques qui complètent les notions générales au moyen d'une série de données de faits, de connaissances concrètes, individuelles. Les sciences idéographiques ne pouvant connaître l'individuel dans toutes ses manifestations, car le problème est sans limites, doivent recourir à un procédé de distribution des objets entre des limites relativement larges de temps et d'espace. « Combien d'objets d'un genre déterminé se comptent dans des limites plus ou moins larges de temps et d'espace », telle est la question fondamentale de la statistique, science idéographique, procédant par catégories. On renonce donc à caractériser les faits isolés, et, au contraire, on les réunit par groupes, ce qui est la méthode d'énumération par catégories des statisticiens.

Nous venons de voir que Sigwart avait aussi trouvé que le caractère essentiel du relevé statistique est l'omission

des caractères individuels ou plutôt leur fusion en de vastes catégories ou totaux.

Une telle explication fournit une notion complète de la statistique méthodologique et présente la statistique comme une application de la logique, coordonnée, non subordonnée à l'induction.

III. — La statistique est-elle une science ou une méthode?

27. Ce que nous avons dit précédemment contient une réponse implicite à cette question, aussi ne faisons-nous plus que résumer brièvement notre point de vue.

La notion de science suppose nécessairement que l'objet étudié est un. Il n'y a pas de science universelle. Ce qui donne à la science son unité, c'est son objet formel. La définition de l'essence d'une chose et de ses propriétés donne naissance à quelques propositions simples, qui conduisent à d'autres, subordonnées aux principes, de sorte que la construction scientifique est tout entière basée sur les principes fournis par l'analyse du sujet (1).

Or, l'objet de la statistique n'est pas un, il est multiple. Il y a la statistique des faits humains et la statistique des faits de l'ordre naturel. Il y a la statistique des naissances, des décès, des mariages, des divorces, des délits et des crimes. Il y a la statistique des quantités d'eau tombée, des jours de gelée, du temps pendant lequel le soleil a brillé. Une science qui s'étendrait à tous ces objets à la fois serait une science monstrueuse, comme Rumelin l'a fait observer avec sa verve habituelle :

« Toute classification des sciences, écrit-il, a pour principe la ressemblance et la différence de leurs objets, non de leurs moyens logiques d'étude. Tout ce que la méthode statistique rassemble chaque jour d'utile pour les diverses bran-

---

(1) Mercier, *Logique*, p. 271.

ches du savoir peut aussi peu rentrer dans le cadre d'une seule science que tout ce que l'induction, l'analogie et l'expérience découvrent chaque jour dans les domaines les plus divers. Ne serait-ce pas une science monstrueuse que celle qui prétendrait embrasser à la fois les isothermes et les isothères, avec les résultats de l'élève du bétail et des remèdes contre la fièvre; les tables de mortalité, la fréquence des meurtres et des suicides avec les bienfaits sociaux des différents systèmes agraires? La statistique ne peut absolument pas être la science de tout ce qui se peut conquérir par la méthode statistique (1). »

28. La science s'efforce de donner une explication des phénomènes. Pourquoi se limiterait-elle à l'emploi, dans ce but, d'une méthode unique? Toute méthode qui peut conduire à la découverte de la vérité doit être employée. Souvent un point spécial du problème qui se pose exige l'emploi de telle méthode; un autre point ne peut être étudié à moins que le savant recoure à une méthode différente de la première : de quel droit interdire l'emploi de telle forme d'investigation scientifique? Or, la statistique, comme science, n'aurait à sa disposition qu'une méthode unique : la méthode statistique. La difficulté serait d'autant plus grande qu'il s'agit ici de l'étude des phénomènes sociaux, qui sont les plus compliqués de tous. Pour trouver l'explication des faits les plus complexes, ceux parmi lesquels on peut démêler l'intervention des causes les plus diverses, on se trouverait limité à l'emploi d'une seule méthode! Sans doute, la méthode statistique est, à l'égard des phénomènes collectifs, une méthode *nécessaire* et non facultative, mais de ce caractère de nécessité on ne peut conclure à celui d'unicité. Or, relisez les définitions de la statistique considérée comme science : « Elle consiste, dit von Mayr, en l'explication des conditions et des phénomènes de la vie

(1) RUMELIN, *loc. cit.*, p. 138.

sociale, explication basée sur une observation numérique en masse, exprimée au moyen de nombres et de mesures. » Il y a donc une confusion entre la science et la méthode, qui ne forment qu'un, alors que l'on considère certains phénomènes ; au contraire, elles se disjoignent lorsqu'on envisage d'autres faits ; conception peu logique, à coup sûr, que celle-là !

29. Aussi, nous paraît-il préférable de considérer la statistique comme une méthode, en réservant le caractère de nécessité de son emploi à l'égard des phénomènes collectifs et plus particulièrement à l'égard de ceux de l'ordre social et de la définir :

« Une méthode qui, par le relevé en masse et l'expression numérique de ses résultats, arrive à la description des phénomènes collectifs et permet de reconnaître ce qu'ils présentent de permanent et de régulier dans leur variété, comme de variable dans leur apparente uniformité. »

Sous cet aspect, la statistique conserve son unité logique. Partie intégrante de la démographie et de la sociologie criminelle, elle fournit à une quantité de recherches scientifiques l'appoint nécessaire de ses constatations numériques et de ses classifications.

Mais la généralité même de ses applications, comme les propriétés fondamentales qui expliquent cette généralité, lui assignent un rang particulier parmi les méthodes générales, à côté de l'induction et de la déduction. Elle est, comme le dit excellemment Lucien March, « une langue commune pour raisonner sur des impressions complexes, toutes les fois que ces impressions ne se fondent point en une apparence homogène ». Au point de vue de son contenu, elle est, comme le dit le même auteur, « la science des faits considérés comme collectivités, la pléthométrie (1) ».

---

(1) LUCIEN MARCH, « Statistique », dans la collection *De la Méthode dans les Sciences*. Paris, Alcan, 1911.

CHAPITRE III

Caractères propres à la statistique

I. — Les caractères de régularité.

30. Devant la complexité des faits qui se présentent à elle dans le domaine des sciences sociales, l'observation commune est impuissante à formuler des conclusions précises. Tous, nous savons que les naissances masculines et féminines se produisent en nombre à peu près égal, car le bon sens nous dit que, s'il n'en était pas ainsi, on verrait, au bout de peu de temps, un sexe représenté par un nombre d'individus beaucoup plus important que l'autre, ce qui, visiblement, n'est pas exact. Mais à cela se borne la présomption que nous pouvons tirer du raisonnement; dans ce cas même, l'observation commune serait de nature à nous induire en erreur, car, dans les limites où elle est capable de s'exercer, elle ne nous révélerait rien autre chose qu'une situation chaotique : ici, nous verrions prédominer les naissances masculines, là ce seraient les naissances féminines qui l'emporteraient. Au total, il serait impossible de rien affirmer de positif.

Au contraire, l'observation scientifique éclaire cette question d'une façon saisissante. Lorsque les observations sont nombreuses, on voit que le rapport entre les naissances masculines et féminines se maintient à peu près exactement dans la proportion de 105-106 garçons pour 100 filles.

La proportion est un peu plus faible en Belgique, mais les chiffres conservent leur portée instructive, quel que soit le rapport.

Pour 100 naissances féminines, nombre de naissances masculines en Belgique :

Périodes.	Naissances masculines. Légitimes.	Illégitimes.
1841 — 1850	105,48	102,54
1851 — 1860	105,44	102,53
1861 — 1870	105,39	103,03
1871 — 1880	104,88	102,37
1881 — 1890	104,63	102,25
1891 — 1900	104,84	102,72
1901 — 1910	104,61	102,71

L'observation scientifique, réalisée par la méthode statistique, montre donc qu'il existe des « régularités » dans un domaine où l'observation commune ne peut rien nous révéler de semblable. Les recherches anthropométriques, qui ont pris tant de développement au cours de la seconde moitié du XIX[e] siècle, ont permis de constater un grand nombre de régularités analogues. Ainsi, la stature comparée des hommes et des femmes a fourni d'intéressants éléments de discussion. Des études statistiques portant sur un nombre élevé d'individus, ont établi que l'homme a une stature plus élevée que la femme à tous les âges, sauf à celui de la puberté, parce que la femme est plus tôt nubile que l'homme ; de même, on a remarqué que la puberté s'accompagne ou est suivie immédiatement d'une forte poussée de croissance, de telle sorte que, selon l'expression d'un auteur, cette période de la vie est semblable à une seconde naissance. Certains auteurs ont soutenu que la tendance à la croissance était constante, mais qu'à partir de certain âge, 35 ans environ, elle était annihilée par différentes circonstances anatomiques et physiologiques. Quoi qu'il en soit, on se trouve en présence de régularités que la méthode statistique est seule à même de faire apparaître ; l'observation commune n'aurait pu fournir à la science les éléments que les recherches anthropométriques ont si largement utilisés.

31. Ce qu'on vient de dire à propos de la stature aux différents âges, on peut le répéter à propos du poids du corps, de l'ampleur thoracique, de la force musculaire, etc. Or, il est intéressant de remarquer que les premiers travaux de Quetelet, ceux qu'il fit paraître entre 1825 et 1835, se rapportent précisément à ces questions (v. *supra* n° 15) et que c'est en se basant sur les constatations qu'il avait faites que le grand statisticien belge prononce pour la première fois le mot de « loi statistique ». « L'homme, écrit-il dès les premières lignes de son *Essai de physique sociale,* publié en 1835, l'homme naît, se développe et meurt d'après certaines lois. »

Il est certain que Quetelet n'a pas entendu donner à ce mot de « loi » le sens rigoureux qu'on y attache quand on parle de lois physiques. Celles-ci expriment un rapport général et constant, mais non absolument nécessaire, en ce sens que la relation entre les phénomènes est établie entre des variables et vérifiée seulement dans les limites d'erreur des expériences. « Les lois qui s'appliquent à l'homme, dit Quetelet, par la manière même dont on les a déterminées, ne présentent plus rien d'individuel : toutes les applications qu'on voudrait en faire à un homme en particulier sont essentiellement fausses ; de même que si l'on prétendait déterminer l'époque à laquelle une personne doit mourir en faisant usage des tables de mortalité (1). » La distinction entre les lois physiques et les lois statistiques appliquées à l'homme est explicite ; toutefois, on ne pourrait affirmer que Quetelet l'ait toujours maintenue aussi nettement ; dans un bon nombre de ses écrits, il est visiblement dominé par la conception mécaniste qu'il avait puisée dans son initiation aux travaux de Laplace et de Fourier. Aussi, ne faut-il pas s'étonner si quelques statisticiens ont, à sa suite, affirmé hardiment qu'il y avait dans les phénomènes

---

(1) QUETELET, *Sur l'homme*, 1835, p. 14.

sociaux comme dans les phénomènes de l'ordre physique, une régularité telle qu'elle pouvait, dans l'un comme dans l'autre cas, être désignée sous le nom commun de « loi ». Dufau n'a-t-il pas écrit que les faits du monde moral ne se comportent pas autrement que ceux du monde physique? « Une étude attentive, ajoute-t-il, montre qu'ils se produisent d'après certaines lois fixes et invariables qui en règlent la succession (1). »

Herschel, qui fut lié d'amitié avec Quetelet, partage les vues de celui-ci et Buckle, fortement inspiré du statisticien belge dans sa conception historique, proclame que les actions humaines au lieu d'être le produit du hasard ou l'effet d'une influence surnaturelle, sont gouvernées par des lois générales en regard desquelles la liberté humaine n'exerce qu'une influence perturbatrice insignifiante. Wagner, à son tour, considère que toute la théorie de la statistique s'appuie sur la conception d'une loi causale générale. C'est la loi universelle des phénomènes qui se succèdent l'un à l'autre et d'après laquelle tout conséquent a un antécédent immuable. La loi du phénomène est le rapport constant entre un phénomène comme effet (ou comme conséquent constant) et un ou plusieurs phénomènes qui en sont la cause (ou antécédents constants). La loi indique l'uniformité du phénomène, elle montre comment la cause domine le phénomène d'une façon toujours égale. Trouver cet état de dépendance d'un phénomène veut dire en trouver la loi.

32. Nous n'entrerons pas, à propos de cette notion de *loi,* dans la discussion, qui s'amène naturellement à la suite, de la question de la liberté et du déterminisme social. L'examen de ce point n'entre pas dans le programme des notions préliminaires que nous traitons brièvement ici, non plus

(1) DUFAU, *De la méthode d'observation dans ses applications aux sciences morales et politiques.* Paris, 1866.

que dans celui de la statistique méthodologique. Aussi, parmi les objections qui furent faites à la conception de « loi », ne prenons-nous que celles qui visent l'essence du terme même et non pas les objections qui impliquent la négation du déterminisme social. Rumelin, avec son esprit critique habituel, a bien formulé l'opinion d'après laquelle le terme de « loi », appliqué aux régularités statistiques, aurait en réalité une portée trop vaste. « La statistique, dit-il, traite la notion de loi, non seulement dans un sens inadmissible pour les autres sciences, mais elle s'imagine même pouvoir construire là-dessus une théorie qui lui est propre. » La « loi des grands nombres », par exemple, est une expression malheureuse aux yeux de Rumelin. Elle éveille l'idée qu'à côté des lois valables pour tous les cas, il en est qui ne commanderaient que les deux tiers ou les trois quarts des cas. « Or, ajoute Rumelin, la généralité est pour tout penseur méthodique le premier et le plus indispensable caractère de la loi. S'il reste un cas où elle est sans effet, bien que sa formule s'y applique, il ne lui reste qu'à conclure que sa formule est fausse. »

Rumelin montre ensuite que la statistique peut conduire à la découverte de lois et que si cette méthode n'est pas la seule qui, dans le domaine des faits sociaux, puisse aboutir à ce résultat, elle est néanmoins une des plus fécondes. Toutefois, dit-il encore, la loi ainsi conquise n'aura plus qu'une forme statistique ou simplement numérique : mais, comme toute autre loi, elle sera générale et sans exception. Elle n'aura plus rien à faire avec le grand nombre, sinon que celui-ci a servi à la découvrir et peut encore aider à la démontrer.

Il ajoute encore : « Il semblerait que nombre de statisticiens, et précisément les fondateurs et les représentants de l'école la plus active et la plus méritante, aient oublié ces échelons naturels d'une étude scientifique, pour attribuer aussitôt ce caractère suprême de loi aux degrés préli-

minaires, aux simples régularités, aux attributs, aux liaisons causales. »

Après avoir rappelé le fait de la prédominance des naissances masculines, Rumelin précise encore sa pensée en disant : « Est-ce à dire qu'il y ait là un ordonnancement divin, comme le pense le premier inventeur de la chose (1), ou même une loi naturelle, comme disent les statisticiens modernes? Ce ne sont, en réalité, que des faits encore inexpliqués pour nous. Ce qui nous manque ici, c'est bien plutôt précisément la loi, et cette loi ne peut être qu'une loi de la généralisation, à découvrir par la physiologie, non par la statistique (2). »

33. A côté de quelques expressions contestables, cette thèse de Rumelin paraît renfermer une grande partie de vérité. L'auteur, par réaction, va trop loin quand il dit d'une manière absolue que la notion de loi entraîne toujours avec elle l'idée d'une liaison constante et générale des phénomènes entre eux. Ce caractère de rigueur absolue que nous sommes enclins à donner aux caractères généraux que nous nommons « lois », est bien plus une abstraction de l'esprit qu'une réalité. Ce qu'on a pris au début pour des lois immuables a été reconnu plus tard, lorsque les recherches méthodiques ont été instituées sur un plan meilleur, comme sujet à certaines variations. Etant données l'infériorité naturelle de nos sens, les limites de notre intelligence, l'imperfection des instruments dont nous nous servons, nous ne pouvons affirmer avec une certitude absolue que les manifestations d'un phénomène que nous considérons comme semblables soient rigoureusement et absolument les mêmes. Personne ne parviendrait à démontrer que deux objets sont exactement semblables, parce qu'on peut concevoir

(1) Sussmilch.

(2) Rumelin, *loc. cit.*, pp. 15 à 19.

l'énumération de leurs qualités individuelles comme atteignant l'infini ou du moins s'étendant à un nombre tel que pratiquement on ne puisse l'épuiser. De même, les manifestations des phénomènes peuvent varier, par des points accessoires que nous n'apercevons pas. Dans notre naïveté ou notre orgueil, nous nous figurons avoir atteint l'absolu.

Au contraire, les sciences expérimentales nous enseignent que beaucoup de manifestations considérées comme uniformes et constantes, sont comprises entre des limites et agissent comme des variables.

Mais Rumelin a raison quand il s'élève contre l'abus qu'on a fait en statistique — et, aurait-il pu ajouter, en économie politique — du mot « loi ». Pour saisir l'erreur de terminologie dans laquelle beaucoup de statisticiens ont versé, il faut s'entendre sur la portée exacte des mots et sur la classification des opérations scientifiques.

Lorsqu'une vérité est connue à l'aide d'un raisonnement déductif qui montre les relations existantes entre deux ou plusieurs phénomènes, le résultat ainsi obtenu se désigne sous le nom de théorème ou encore d'hypothèse scientifique. Les vérités démontrées par la géométrie ne sont pas des lois, mais des théorèmes. On ne dit pas la loi du carré de l'hypoténuse, mais le théorème du carré de l'hypoténuse.

Au contraire, si la même vérité ne pouvait être atteinte que par une longue suite d'expérimentations ou d'observations, au lieu d'être déduite par le raisonnement, on énoncerait un principe. Si entre les faits, on établissait une relation causale, on aurait énoncé une loi.

34. Benini fait remarquer avec raison qu'il importe de distinguer entre les « lois » et les « notions empiriques concernant l'uniformité et la régularité des phénomènes ». Savoir que telle ou telle année il y aura des éclipses du soleil ou de la lune est une notion simple; savoir que, à chaque cycle de 223 lunaisons (18 ans, 11 jours, 7-8 heures),

on a une répétition assez approximative des éclipses précédentes, signifie qu'on a la notion empirique d'une uniformité naturelle; déterminer enfin les antécédents de cette périodicité, les rapporter aux lois de la mécanique céleste, est affirmer une loi.

Or, en statistique, on a souvent attribué avec prodigalité le nom de lois à de simples régularités ou uniformités empiriques, sans que l'on ait réuni la moindre donnée explicative relativement aux causes qui provoquent cette régularité ou aux circonstances qui tantôt en permettent la manifestation, tantôt l'empêchent d'apparaître. Supposons qu'à l'aide de recherches physiologiques, on ait découvert, comme on l'a prétendu déjà, que tel régime alimentaire appliqué aux parents aurait pour conséquence de provoquer la naissance d'enfants mâles; que l'âge des époux étant dans tel rapport, on aurait chance de voir des naissances féminines se produire en plus grand nombre que des masculines. Supposons encore que ces faits aient été vérifiés un très grand nombre de fois, qu'en y appliquant les méthodes inductives on ait reconnu entre eux un lien causal indiscutable, alors on serait en droit de donner le nom de lois à des constatations de l'espèce.

35. Mais ces conditions ne sont pas réalisées dans la méthode statistique. Ce que la statistique nous livre, c'est une constatation obtenue à l'aide d'une méthode scientifique, sans plus. Elle ne nous donne aucune explication directe. Elle ne fait immédiatement ressortir aucun lien causal entre les phénomènes. Le statisticien nous dit : J'ai observé un grand nombre de naissances au point de vue de la répartition des sexes. Je puis vous affirmer que les naissances masculines sont un peu plus nombreuses que les naissances féminines, et cela dans une proportion de 4-5 ou 5-6 p. c.

La statistique ne nous dit pas : cet excédent des naissances masculines est dû à telle et telle cause. Elle constate le fait. Elle énonce une régularité de la nature, mais elle ne

va pas plus loin. C'est ajouter au sens naturel des mots que de voir dans cette constatation une loi. Nous connaîtrons la loi par des recherches successives, si nous arrivons jamais à la connaître. Ce sera peut-être l'œuvre d'un physiologiste, ou d'un embryologiste, mais il n'est pas à présumer que ce soit celle d'un statisticien. Peut-être, dans les recherches qui conduiraient à la découverte de la loi, y aurait-il lieu de recourir à la méthode statistique. La chose est vraisemblable, mais elle ne fait que montrer que la statistique est avant tout une méthode scientifique qui trouve sa place dans des sciences distinctes. En résumé, la statistique constate des régularités naturelles et ne découvre pas de lois. En cela elle remplit sa fonction de méthode scientifique et le rôle qu'elle joue à cet égard est assez utile, assez glorieux pourrait-on dire, pour qu'il ne soit pas besoin, pour en renforcer la valeur, de détourner les mots de leur acception véritable et de nommer loi la constatation des régularités naturelles par le procédé statistique.

S'il fallait encore ajouter quelque chose aux arguments qui viennent d'être exposés, il suffirait, pour emporter la conviction, de faire remarquer que les tendances et les régularités décorées du nom de « lois statistiques » n'ont rien d'absolu, ni de permanent. Absolues, ces « lois » ne le sont pas dans le temps, puisqu'elles sont sujettes à modifications et qu'elles peuvent même cesser de se manifester; elles n'ont rien d'absolu dans l'espace puisque leurs manifestations varient d'un pays à un autre, même d'une province à une autre; elles ne sont même pas absolues à un moment donné, dans un lieu donné, puisque, pour la plupart, elles varient dans leur manière d'être selon qu'elles se manifestent dans des classes sociales différentes. Les constances statistiques sont purement relatives; résultant d'une action combinée d'éléments constants et variables, elles subissent les influences du temps et du lieu où on les observe; vraies, à un moment, pour un groupe; elles ne le sont pas pour les individus qui composent ce groupe. Ce sont, comme dit Ga-

baglio (1), des lois de fait, secondaires, dérivées, empiriques, comme les causes desquelles elles résultent.

II. — Notions générales sur les combinaisons et les probabilités.

36. La connaissance scientifique dans son expression la plus haute pourrait se caractériser par un mot : la certitude, mais ce sommet, que nous entrevoyons dans nos rêves, n'est pas près d'être atteint. Toute connaissance est relative ; la plus complète ne fait encore que réunir un certain nombre — élevé, il est vrai, — de probabilités. N'arrive-t-il pas que les théories les plus généralement admises sont déracinées par des découvertes nouvelles ? C'est une humiliation que nous subissons avec patience ; elle n'empêche pas que, sur ces données nouvelles, nous construisions une théorie neuve, car le besoin de synthétiser en un corps de doctrine les expériences journalières est universel et impérissable. La conviction que le monde n'est pas régi par des lois aveugles, l'espoir de découvrir la raison des phénomènes au milieu desquels nous vivons, est un ressort qui ne s'use pas. Comme le dit si bien M. Emile Borel : « Le jour où l'homme a compris qu'il pouvait se proposer un tel but, il ne s'en est jamais laissé détourner ; même dans les périodes les plus sombres de l'histoire, lorsque les soucis de la vie matérielle absorbaient presque toutes les énergies, il s'est trouvé des chevaliers servants de la raison pour conserver et transmettre le flambeau de la pensée antique (2). »

L'expérience isolée ne peut satisfaire notre soif de savoir. Nous voulons connaître la façon dont les phénomènes s'enchaînent et s'ordonnent. Rechercher les causes des phénomènes est la tâche la plus essentielle qui s'impose à nous,

---

(1) GABAGLIO, *loc. cit*., p. 384.
(2) EMILE BOREL, *Le Hasard*, p. 4. Paris, Alcan, 1914.

car si nous connaissions, par impossible, les causes de tous les phénomènes, nous connaîtrions, du même coup, toutes les circonstances de leur répétition éventuelle, en quelque sorte nous connaîtrions l'avenir. Il importe peu que la raison nous dise qu'un tel rêve est vain. Il y a une telle distance entre cette connaissance générale et absolue et la relativité autant que la petitesse de notre savoir, qu'on peut toujours essayer de gravir un échelon. D'ailleurs, un progrès en appelle un autre, aucune connaissance positive ajoutée à notre bagage intellectuel n'est sans utilité pour la conquête d'une notion nouvelle qui, auparavant, nous paraissait bien chimérique. C'est une loi de la science qu'une connaissance nouvelle en appelle invinciblement une autre. Pas à pas, nous nous acheminons vers le but : qu'importe qu'il soit encore lointain, si nous nous en rapprochons?

37. Nous avons dit plus haut ce qu'il fallait entendre par « loi statistique » et nous avons montré qu'il s'agissait de désigner sous ce nom des régularités constatées parmi les phénomènes. L'explication causale est absente de cette conception, mais rien que la découverte de la régularité est précieuse parce qu'elle met le penseur sur la piste de l'explication causale. A l'aide des méthodes d'analyse, le chercheur parviendra peut-être à isoler les causes constantes des causes accidentelles et alors il se trouvera sur le chemin qui conduit à l'explication finale; même réduite à ces proportions relativement modestes, la recherche scientifique n'en renferme pas moins une difficulté redoutable.

La complexité du monde et de la vie est inimaginable. Elle fait d'ailleurs partie de leur nature intime. L'immobilité complète n'évoque chez nous que l'idée de la mort. La vie, c'est le mouvement, c'est le changement, c'est la variété, c'est la complexité. Les causes s'enchaînent et s'entre-croisent, elles se multiplient par leurs actions réciproques. Comment faire pour les classer, pour les dénombrer? Car avant de connaître les phénomènes probables

nous devons prendre conscience des phénomènes possibles. C'est la raison pour laquelle, avant d'esquisser la notion de la « probabilité statistique », nous devons aborder en peu de mots les éléments du calcul des combinaisons et des permutations.

38. Entre plusieurs objets que nous observons, il peut exister des arrangements, se présenter des changements, s'imaginer des combinaisons extraordinairement nombreuses et variées.

On appelle « analyse combinatoire » cette partie des mathématiques qui a pour but d'énumérer les différentes manières dont on peut ranger des objets donnés dans des circonstances déterminées. Les cas les plus simples de l'analyse combinatoire consistent en ce qu'on appelle : A, les arrangements ; B, les permutations ; C, les combinaisons. Comme le fait justement observer Stanley Jevons, le calcul combinatoire est une des parties les plus curieuses des mathématiques ; les formes des expressions algébriques sont déterminées par les principes des combinaisons, et les opérations fondamentales du calcul des probabilités ne diffèrent pas de celles de l'analyse combinatoire. Les nombres pour ainsi dire infinis, en tout cas inexprimables, auxquels on arrive rapidement dans les permutations, ont été de tout temps un objet de curiosité et d'étonnement. Pascal et Leibniz se sont appliqués avec passion au calcul des permutations et des combinaisons, qui a été également un sujet d'études auquel Jacques Bernouilli consacra une partie de son existence. Dans les lignes qui suivent nous ne nous occupons que des arrangements, permutations et combinaisons sans répétition. Aussi bien il existe des arrangements, permutations et combinaisons avec répétition, c'est-à-dire dans lesquels les mêmes objets sont répétés identiquement.

39. On appelle *arrangements* les modifications que l'on fait subir à un certain nombre d'objets, au point de vue de

leur ordonnancement, de manière à former des groupes de n à n objets parmi m objets donnés. Le symbole sous lequel on désigne les arrangements est A m n et la formule du calcul est :

$$A\,m\,n = m\,(m-1)\,(m-2)\,(m-3)\,\ldots\ldots\,(m-n+1), \qquad (1)$$

formule qui peut se traduire par la règle : le nombre des éléments à disposer est multiplié successivement par ce nombre diminué chaque fois d'une unité et cela autant de fois qu'il y a d'unités de la classe. Il faut et il suffit qu'aucun groupe ainsi formé ne soit identique à un autre, soit par la nature des éléments qu'il comprend, soit par l'ordre dans lequel ces éléments sont énumérés. Ainsi : D E F, lettres prises sur un ensemble de 6 lettres, constituent un groupe différent de A B C, mais C B A forment aussi un groupe différent de A B C.

Soit l'exemple suivant : une maîtresse de maison recevant chez elle vingt invités, veut, après une soirée musicale, leur offrir à souper par « petites tables » de quatre personnes chacune. Combien d'arrangements pourra-t-elle imaginer pour faire varier la composition des tables et l'ordre dans lequel les vingt invités pourront se placer? Nous avons :

$$A\,20{,}4 = 20\,.\,19\,.\,18\,.\,17 = 85{,}680. \quad 116\,280$$

Supposons maintenant que quatre personnes de la maison prennent place avec les vingt invités et que les arrangements possibles soient calculés sur l'ensemble de 24 personnes, on a :

$$A\,24{,}4 = 24\,.\,23\,.\,22\,.\,21 = 234{,}024.$$

On voit par là que le fait d'ajouter un petit nombre d'unités à un nombre primitivement choisi a pour conséquence d'augmenter prodigieusement le nombre d'arrangements possibles.

Ainsi, si au lieu de 24 personnes, nous prenons pour exemple les 26 lettres de l'alphabet rangées quatre par quatre, nous avons :

$$A\,26{,}4 = 26 \,.\, 25 \,.\, 24 \,.\, 23 = 358{,}800.$$

40. Si le nombre d'arrangements divers que l'on peut obtenir entre un nombre relativement peu élevé de choses différentes est un sujet d'étonnement, que faudrait-il dire des nombres auxquels on arrive dans le calcul des permutations?

Le symbole des permutations est P n et la formule du calcul à opérer est :

$$P\,n = 1 \,.\, 2 \,.\, 3 \,.\, \ldots \,.\, n. \qquad (2)$$

On donne le nom de *permutations* au nombre de manières dont on peut ranger n objets, ou encore au nombre de groupes différents que l'on peut former en les rangeant successivement, dans des ordres différents.

Dans les permutations les groupements constitués contiennent tous les mêmes éléments, mais chaque groupement est considéré comme différent d'un autre du moment que les éléments qu'il contient sont rangés dans un ordre qui n'a pas encore été réalisé. Le nombre de n éléments est donc égal au produit des n nombres entiers. Remplaçons n par un chiffre, nous avons :

$$P\,8 = 1 \,.\, 2 \,.\, 3 \,.\, 4 \,.\, 5 \,.\, 6 \,.\, 7 \,.\, 8 = 40{,}320.$$

Supposons que nous ayons à présenter à un haut personnage dix personnes entre lesquelles ne se marque aucune distinction essentielle, de façon que l'ordre de présentation puisse être indifféremment modifié. Nous avons :

$$P\,10 = 1 \,.\, 2 \,.\, 3 \,.\, 4 \,.\, 5 \,.\, 6 \,.\, 7 \,.\, 8 \,.\, 9 \,.\, 10 = 3{,}628{,}800.$$

Si le nombre de ces personnes atteignait le chiffre de 12, nous atteindrions le chiffre de 479 millions.

On appelle *factorielle* le produit de facteurs en progression arithmétique, donc le produit de tous les nombres entiers depuis l'unité jusqu'à la valeur n exprimée dans le symbole P n. Les factorielles ont été calculées pour un certain nombre de chiffres. Voici celles qui comprennent les 12 premiers nombres :

$$24 = \underline{|4} \quad (1 . 2 . 3 . 4)$$
$$120 = \underline{|5} \quad (1 . 2 . 3 . 4 . 5)$$
$$720 = \underline{|6} \quad (1 . 2 . 3 . 4 . 5 . 6)$$
$$5.040 = \underline{|7} \quad (1 . 2 . 3 . 4 . 5 . 6 . 7).$$
$$40.320 = \underline{|8}$$
$$362.880 = \underline{|9}$$
$$3.628.800 = \underline{|10}$$
$$39.916.800 = \underline{|11}$$
$$479.001.600 = \underline{|12}$$

41. Il importe de ne pas confondre les permutations dont nous venons de parler et les combinaisons auxquelles est consacré le présent numéro. Il ne faut pas confondre non plus les combinaisons avec les arrangements. A cette fin, reprenons l'exemple que nous avons donné sous le n° 39 en parlant des arrangements.

Une maîtresse de maison, avons-nous dit, ayant à placer vingt invités, par « petites tables » de quatre personnes et voulant, à la fois, déterminer la composition de chaque table et la place que chacun y occupera, a le choix entre 85,680 façons différentes de ranger son monde.

Le procédé auquel elle a eu recours pour déterminer ce chiffre est celui donné par la formule des arrangements :

$$\text{A } 20,4 = 20 \ . \ 19 \ . \ 18 \ . \ 17 = 85,680.$$

Mais l'énormité même du nombre de manières dont il est possible de disposer les vingt invités constitue une difficulté pratique. La maîtresse de maison se préoccupe alors de simplifier les données du problème et se dit qu'il serait

bien suffisant que la composition des tables soit arrêtée à l'avance sans qu'on se préoccupe de fixer la place que chacun des quatre invités occupera à cette table : nous sommes ici dans le cas des combinaisons.

On appelle *combinaisons* de m objets pris n à n les groupes que l'on peut former en prenant n objets parmi ces m objets donnés, de manière à ce que deux groupes diffèrent par *la nature* des objets qui servent à les former. Dans les arrangements, *la place* des objets intervient en même temps que leur nature. On peut encore dire que « les combinaisons sont des groupements que l'on peut former entre plusieurs (m) éléments, pris deux à deux, trois à trois, ou en un nombre quelconque ne dépassant pas n, à condition que chaque groupe contienne au moins un élément autre que dans les autres groupes.

$$Cm, n = \frac{A\, m.n}{P\, n} \text{ ou } \frac{m(m-1)\ldots.(m-n+1)}{1\,.\,2\,.\,3\ldots\ldots n} \qquad (3)$$

Le problème que notre maîtresse de maison a à résoudre est donc le suivant :

$$C.\ 20{,}4 = \frac{20\,.\,19\,.\,18\,.\,17}{1\,.\,2\,.\,3\,.\,4} = \frac{85.680}{24} = 3570.$$

Il y a donc 3,570 combinaisons de table différentes au lieu de 85,680 comme dans le cas précédent, ce qui est encore bien suffisant à créer des maux de tête à notre maîtresse de maison ; aussi est-il probable qu'elle se résoudra à laisser ses invités se grouper eux-mêmes d'après leurs goûts et leurs sympathies, et ce sera pour un mieux.

Si, au lieu de faire des tables de quatre, on arrangeait des tables de trois, il n'y aurait que 840 combinaisons possibles et si l'on voulait organiser des tête-à-tête, il n'y en aurait que 140 possibles, car :

$$C.\ 20{,}3 = \frac{20\,.\,19\,.\,18}{1\,.\,2\,.\,3} = \frac{5040}{6} = 840$$

ou

$$C.\ 20{,}2 = \frac{20\,.\,19}{1\,.\,2} = \frac{280}{2} = 140.$$

42. Les différences numériques qui se remarquent dans les résultats des arrangements, permutations et combinaisons sont considérables. La raison de ces différences est facile à apercevoir. Dans les permutations, nous envisageons tous les éléments et tous les changements possibles entre ces éléments ; les permutations sont des énumérations complètes de tous les cas possibles, chaque cas ne différant d'un autre que par la place qu'il occupe. C'est donc comme si l'on écrivait toute la série autant de fois qu'on pourrait le faire à la condition unique qu'aucune des énumérations ne soit rigoureusement identique à une précédente. On arrive de la sorte, et très rapidement, à des nombres incommensurables parce que chaque produit partiel est multiplié par un facteur supérieur d'une unité au précédent. Nous avons dit que la factorielle de 12 est déjà 479,001,600. Les logarithmes des factorielles jusqu'à 265 ont été publiés par une association scientifique anglaise, la « Society for the diffusion of the Useful Knowledge » ; pour exprimer en chiffres une factorielle de cette importance, il faudrait utiliser 529 chiffres.

Dans les arrangements, au contraire, l'accroissement des données numériques est limité par le nombre des choses que l'on considère en même temps, soit deux à deux, trois à trois, quatre à quatre. Partant du nombre le plus élevé, on diminue les facteurs successifs chaque fois d'une unité. On atteindrait des nombres aussi élevés que dans les permutations dans une hypothèse seulement : celle où l'on arrangerait les objets considérés sous un nombre d'unités de la classe égal au nombre total des termes, mais alors on serait dans le cas des permutations et non plus des arrangements.

Toute combinaison est, dans sa nature intime, toujours différente d'une autre combinaison, car un des éléments qui en faisait partie cède sa place à un autre : nous avons déjà dit que sont exclus les groupes qui contiennent les mêmes éléments mais en ordre divers, hypothèse qui fait partie, au contraire, des arrangements et des permutations.

Il est donc naturel que les combinaisons soient le mode d'énumération qui contient les éléments les moins nombreux.

On a aussi fait remarquer que l'ajoute d'un seul élément aux arrangements, permutations et combinaisons augmente les nombres dans une proportion bien plus élevée qu'on ne pourrait s'y attendre à première vue. Ce fait résulte de l'addition d'un facteur nouveau qui multiplie tous les autres.

Dans les permutations, le dernier facteur est toujours supérieur d'une unité au précédent.

Ainsi, nous avons :

$$P\,n\,6 = 1\,.\,2\,.\,3\,.\,4\,.\,5\,.\,6 = 720$$

et

$$P\,n\,7 = 1\,.\,2\,.\,3\,.\,4\,.\,5\,.\,6\,.\,7 = 5040.$$

Dans les arrangements, chaque facteur nouveau diminue d'une unité seulement par rapport au précédent et il multiplie tous les produits antérieurs à son introduction.

$$A\,90{,}3 = 90\,.\,89\,.\,88 = 704{,}880.$$
$$A\,90{,}4 = 90\,.\,89\,.\,88\,.\,87 = 61{,}324{,}560.$$

La somme obtenue dans le cas de A 90, 4 est donc 87 fois plus élevée que dans le cas de A 90, 3.

Il suffit de se reporter à ces exemples pour se faire une idée de la complexité des phénomènes naturels dans lesquels plusieurs facteurs se combinent.

43. La construction du triangle arithmétique, appelé aussi triangle de Pascal, permet de rechercher immédiatement quel est, pour un nombre déterminé d'objets, le nombre de combinaisons possibles et quel est le nombre de manières dont *m* choses peuvent être choisies en combinaison de *n* choses données. La solution de ce problème nous permet en même temps de résoudre un certain nombre de questions simples du calcul des probabilités. Stanley Jevons,

dans son ouvrage cité plus haut, donne la formule suivante pour la construction du triangle arithmétique :

On commence, dit-il, par inscrire l'unité au sommet, puis, à la ligne suivante, nous plaçons une seconde unité à la droite de la première. Pour obtenir la troisième ligne de chiffres, nous reculons la ligne précédente d'un rang à droite et nous les additionnons aux mêmes chiffres, tels qu'ils étaient avant d'être reculés. Soit l'exemple suivant :

1°	1
2°	. 1
	1 . 1
3°	1 . 2 . 1
	1 . 2 . 1
4°	1 . 3 . 3 . 1
	1 . 3 . 3 . 1
5°	1 . 4 . 6 . 4 . 1
	1 . 4 . 6 . 4 . 1
6°	1 . 5 . 10 . 10 . 5 . 1
	1 . 5 . 10 . 10 . 5 . 1
7°	1 . 6 . 15 . 20 . 15 . 6 . 1

La règle est formulée d'une façon un peu différente par M. Em. Borel :

« On inscrit sur une première ligne le nombre 1 répété deux fois, puis on calcule chaque nombre à inscrire dans les lignes suivantes, en ajoutant au nombre inscrit immédiatement au-dessus de lui le nombre inscrit à la gauche de celui auquel on l'ajoute; dans l'application de cette règle, on suppose mentalement que les lignes sont prolongées à droite et à gauche par des zéros (1). »

Nous reproduisons ici le triangle arithmétique calculé pour les 17 premières lignes. Il est évident que ce calcul peut, en théorie, être répété sans limite, mais on arrive rapidement à des chiffres tellement forts que, pratiquement, la solution possible est renfermée dans des limites relativement étroites.

---

(1) Emile Borel, *Eléments de la théorie des probabilités*. Paris, Hermann, 1909, p. 9.

LIGNES	Colonnes																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
1	1	—																1
2	1	1	—															2
3	1	2	1	—														4
4	1	3	3	1	—													8
5	1	4	6	4	1	—												16
6	1	5	10	10	5	1	—											32
7	1	6	15	20	15	6	1	—										64
8	1	7	21	35	35	21	7	1	—									128
9	1	8	28	56	70	56	28	8	1	—								256
10	1	9	36	84	126	126	84	36	9	1	—							512
11	1	10	45	120	210	252	210	120	45	10	1	—						1024
12	1	11	55	165	330	462	462	330	165	55	11	1						2048
13	1	12	66	220	495	792	924	792	495	220	66	12	1	—				4096
14	1	13	78	286	715	1287	1716	1716	1287	715	286	78	13	1	—			8192
15	1	14	91	364	1001	2002	3003	3432	3003	2002	1001	364	91	14	1	—		16384
16	1	15	105	455	1365	3003	5005	6435	6435	5005	3003	1365	455	105	15	1	—	32768
17	1	16	120	560	1820	4368	8008	11440	12870	11440	8008	4368	1820	560	120	16	1	65536

44. Le triangle arithmétique réunit une série de propriétés. En premier lieu, nous remarquons que le total des nombres inscrits sur chaque ligne augmente selon la progression 2, 4, 8, 16, 32, 64, etc., en sorte que la somme de chaque ligne horizontale représente le double de la ligne immédiatement supérieure et la moitié de la ligne subséquente.

Il suit également du mode de construction du tableau que la différence des nombres dans chaque colonne peut être trouvée dans la colonne précédente à gauche, un rang plus haut. L'unité qui apparaît dans la première colonne est la première différence des nombres inscrits dans la seconde colonne et ainsi de suite.

La disposition des nombres est symétrique, c'est-à-dire qu'ils vont en augmentant jusqu'au milieu, puis ils diminuent de la même façon. Dans les lignes de rang pair, le partage se fait exactement au milieu, en un nombre unique; dans les lignes de rang impair, ce partage égal ne peut être réalisé et l'on trouve le nombre le plus élevé répété deux fois. A chaque ligne, le point le plus élevé est donc tantôt indiqué une seule fois, tantôt deux fois, et la place que ce nombre occupe est chaque fois d'une rangée de plus à droite.

Enfin, le triangle arithmétique permet de trouver immédiatement les coefficients des puissances successives d'un binôme : la première colonne horizontale renferme les coefficients de la première puissance du binôme; la deuxième, les coefficients de la deuxième puissance; la n^{me}, les coefficients de la n^{me} puissance. Le triangle arithmétique facilite ainsi l'application du binôme de Newton, puisque nous savons d'ailleurs que les exposants de la lettre a vont sans cesse en diminuant et que ceux de la lettre b vont en augmentant.

Le triangle arithmétique présente le grand avantage de résoudre immédiatement les problèmes de combinaison, en indiquant de combien de manières on peut choisir m choses en combinaison de n choses. Voici un exemple emprunté à Stanley Jevons : on demande de combien de manières une sous-commission de 5 membres pourra être choisie parmi un comité qui compte 9 membres.

Le calcul donne :

$$9, 5 = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 126,$$

mais nous pouvons supprimer tout calcul, en nous servant du triangle arithmétique; il suffit de nous reporter à la

dixième ligne et de consulter le chiffre inscrit à la sixième colonne, où nous trouvons 126 (1).

Parmi 10 invités, on désire faire des classements 3 par 3. Combien y a-t-il de classements possibles? Nous consultons la onzième ligne, quatrième colonne, et nous lisons la réponse 120. En effet, le calcul donne :

$$C\ 10,\ 3 = \frac{10\ .\ 9\ .\ 8}{1\ .\ 2\ .\ 3} = 120.$$

Nous empruntons encore à M. Borel le problème suivant dont la solution est donnée immédiatement par la simple lecture du triangle arithmétique :

« Y a-t-il plus d'avantage à parier que l'on gagnera au jeu de pile ou face, au moins 5 parties sur 7 ou que l'on en gagnera au moins 6 sur 9 ? »

Nous reportant au triangle arithmétique construit d'après la méthode de Stanley Jevons, nous avons à consulter, pour déterminer la valeur du premier cas 5/7, la huitième ligne, sixième colonne et, bien entendu, à additionner les nombres consécutifs dont la somme formera le total des cas favorables. Nous lisons les nombres suivants :

$$21 + 7 + 1 = \frac{29}{128}.$$

En procédant de même pour la seconde hypothèse, nous avons à la dixième ligne, septième colonne :

$$84 + 36 + 1 = \frac{130}{512}.$$

La première fraction représente : $\frac{1}{4} - \frac{3}{128}$; la seconde vaut $\frac{1}{4} + \frac{1}{256}$; la seconde hypothèse est donc légèrement plus favorable que la première.

(1) Ceci d'après la règle de formation du triangle donnée par Stanley Jevons. Si l'on adopte la règle indiquée par Borel, il faut consulter la neuvième ligne.

45. La formule mathématique des probabilités est la même que celle des combinaisons, d'où il suit que le triangle arithmétique est également capable de fournir la réponse à donner aux questions simples du calcul des probabilités, comme nous venons de le voir par l'exemple des paris engagés sur la chance respective de gagner 5 parties sur 7 ou 6 parties sur 9.

La notion la plus simple de la probabilité mathématique s'exprime en disant que cette probabilité est mesurée par le rapport entre le nombre de cas favorables et le nombre de cas possibles. Les cas possibles sont mesurés par le nombre de combinaisons possibles ; les cas favorables sont ceux dont nous attendons l'arrivée. La probabilité la plus grande est exprimée par le rapport le plus élevé existant entre les deux nombres de cas possibles et de cas favorables. En dehors de ce point, la probabilité décroît graduellement.

La théorie des probabilités dans ses rapports avec la statistique sera examinée en détail à l'endroit où l'on parlera de la loi de la répartition des erreurs accidentelles. Pour le moment, il suffit de rappeler que les phénomènes collectifs comportent une série de manifestations indépendantes parmi lesquelles nous avons à rechercher quelle est la plus probable. Or, pour exprimer la probabilité de plusieurs événements qui sont indépendants entre eux, il faut considérer deux hypothèses : s'il y a deux événements ou plus qui ne peuvent arriver en même temps, la probabilité que l'un ou l'autre arrivera est la *somme* de leurs probabilités respectives ; s'il y a deux événements dépendant l'un de l'autre de telle sorte que le premier doive avoir lieu avant que le second puisse se produire, la probabilité de l'arrivée simultanée des deux événements est le *produit* de la probabilité du premier par la probabilité qu'après l'arrivée du premier le second surviendra. La probabilité composée d'événements indépendants, observée par des épreuves ou observations successives, présente un développement cor-

respondant, dans sa forme mathématique, à celui d'un binôme élevé à une puissance égale au nombre des épreuves ou observations. On connaît la forme du binôme qui, exprimée graphiquement, correspond à la courbe binomiale si souvent utilisée par Quetelet et les autres statisticiens (1).

Le fait que l'on connaît la marche suivie par les probabilités dans leur ordre de succession permet de rechercher les combinaisons les plus probables parmi les événements composés; la probabilité des différentes combinaisons des éléments de la série correspond donc aux termes de ce développement et le terme qui a la valeur la plus élevée indiquera la *combinaison qui se présente comme la plus favorable* (2).

Théoriquement, les phénomènes collectifs se présenteraient comme des probabilités composées et l'on pourrait déterminer le nombre de leurs combinaisons possibles et l'ordre probable de leur arrivée.

46. Si nombreuses sont les régularités d'ordre statistique constatées à propos des phénomènes démographiques et moraux, que leur multiplicité même est un argument contre le caractère de « loi » que de nombreux statisticiens ont voulu leur attribuer. Les lois scientifiques sont d'une rare et pure essence et on ne conçoit pas qu'à l'aide de quelques chiffres on puisse en déterminer pour ainsi dire à volonté, pas plus à propos des faits économiques et sociaux qu'à propos des phénomènes physiques, astronomiques ou

---

(1) Si l'on donne à *m*, dans la formule binominale (a+b) *m*, la valeur successive de 2, 3, 4, 5, 6, on a :

$$(a+b)^2 = a^2 + 2\,ab + b^2.$$
$$(a+b)^3 = a^3 + 3\,a^2b + 3\,ab^2 + b^3.$$
$$(a+b)^4 = a^4 + 4\,a^3b + 6\,a^2b^2 + 4\,ab^3 + b^4.$$
$$(a+b)^5 = a^5 + 5\,a^4b + 10\,a^3b^2 + 10\,a^2b^3 + 5\,ab^4 + b^5.$$
$$(a+b)^6 = a^6 + 6\,a^5b + 15\,a^4b^2 + 20\,a^3b^3 + 15\,a^2b^4 + 6\,ab^5 + b^6.$$

(2) Bosco, *loc. cit.*, p. 117. Pour la loi des coefficients et la loi des exposants, voir les explications données *supra*, n° 35.

biologiques. Les Newton sont rares. Il n'appartient pas à la statistique de revendiquer pour elle une fécondité aussi prodigieuse dans la découverte des « lois ». Ce que la statistique constate donc, ce sont bien des « régularités », des tendances générales résultant de causes multiples, les unes générales, les autres accidentelles, tendances qui perdurent et dont les manifestations se reproduisent, mais sans rien d'absolu et de définitif.

La question se pose tout naturellement de savoir si l'application du calcul des probabilités aux faits relevés par la statistique est possible, légitime et utile. Nous savons que certains phénomènes se sont produits dans telles conditions et en tel nombre; nous prévoyons aussi qu'ils se reproduiront; peut-on, à l'aide du calcul, déterminer la mesure de leur arrivée éventuelle? Le calcul des probabilités appliqué aux phénomènes statistiques présente une réelle utilité, ainsi que nous le montrons plus loin par quelques exemples. Il faut distinguer deux espèces de probabilités : la probabilité *a priori* et la probabilité *a posteriori*. La seconde est le résultat de l'expérience; après avoir observé et dénombré un grand nombre de faits, après avoir constaté qu'ils se représentent d'une façon régulière, on conclut à la répétition du phénomène dans les mêmes conditions. Mais la probabilité, au sens rigoureux du mot, est donnée par la probabilité *a priori* qui résulte du calcul. On peut la définir — en se bornant à la notion de la probabilité simple — « le rapport entre le nombre de cas favorables à l'attente d'un événement et le nombre de cas possibles à la condition que tous les cas, favorables ou non, soient ou puissent raisonnablement être supposés possibles au même degré (1) ».

La portée générale du calcul des probabilités a été ainsi résumée par M. Poincaré : « La méthode des sciences physiques, écrivait cet éminent mathématicien, repose sur l'in-

(1) Benini, *Statistica metodologica*, p. 213.

duction qui nous fait attendre la répétition d'un phénomène quand se reproduisent les circonstances où il avait une première fois pris naissance. Si *toutes* ces circonstances pouvaient se reproduire à la fois, ce principe pourrait être appliqué sans crainte; mais cela n'arrivera jamais, quelques-unes de ces circonstances feront toujours défaut. Sommes-nous absolument sûrs qu'elles sont sans importance? Cela pourra être vraisemblable; cela ne pourra pas être rigoureusement certain. De là, le rôle considérable que joue dans les sciences physiques la notion de probabilité (1). »

Il n'y a pas de doute que de nombreux phénomènes collectifs se comportent dans leurs combinaison et distribution comme l'indique le calcul des probabilités. Mais la ressemblance entre les données résultant du calcul et les données provenant de l'expérience n'est pas parfaite. « Il ne faut pas perdre de vue, écrit M. Borel, qu'il y a la même différence entre ces probabilités statistiques et les probabilités abstraitement et rigoureusement définies qu'entre les figures étudiées en géométrie et les représentations plus ou moins grossières qu'on rencontre dans la nature; entre une sphère, par exemple, et une orange (2). »

47. L'application du calcul des probabilités aux données de la statistique présente cet avantage d'établir précisément quelle est la différence entre l'orange et la sphère tracée par le géomètre; en comparant les résultats de l'expérience avec ceux du calcul on aura la mesure et la direction des déviations provoquées par un complexus de causes agissant sur le phénomène naturel dans les conditions où il se produit.

---

(1) H. POINCARÉ, *La Science et l'Hypothèse* Paris, Flammarion, « Introduction », p. 6.

(2) EMILE BOREL, *Eléments de la théorie des probabilités*. Paris, Hermann, p. 158.

La différence observée entre le résultat théorique de la probabilité mathématique et le résultat pratique constaté par le relevé statistique est plus ou moins grande, selon qu'il s'agit de faits de la nature ou de faits influencés par la volonté libre de l'homme. Les faits de la nature sont doués d'une plus grande fixité et s'écartent moins du calcul; on les rencontre surtout en démographie; les faits influencés par la volonté de l'homme se trouvent surtout dans le domaine de la statistique morale, mais il en est également en démographie, par exemple les combinaisons matrimoniales étudiées sous le rapport de l'état civil des époux ou au point de vue de leur instruction. Quand on parle du caractère de fixité que présentent les faits de la nature, on se place à un point de vue relatif, car les phénomènes les plus constants dans leurs manifestations quantitatives présentent à la longue d'assez curieuses modifications. Ainsi en est-il du rapport bien connu entre le nombre de naissances masculines et celui des naissances féminines, comme le montre l'exemple suivant :

Proportion des naissances masculines pour cent naissances féminines.

Années	Belgique	Années	Italie (1)
1841 à 1850 . . .	105.48	—	—
1851 à 1860 . . .	105.44	—	—
1861 à 1870 . . .	105.39	—	—
1871 à 1880 . . .	104.88	1872 à 1875 . . .	106,48
—	—	1876 à 1880 . . .	106,34
1881 à 1890 . . .	104,63	1881 à 1885 . . .	106,06
—	—	1886 à 1890 . . .	105,85
1891 à 1900 . . .	104.84	1891 à 1895 . . .	105,73
—	—	1896 à 1902 . . .	105.67
1901 à 1910 . . .	104,61	—	—

Les chiffres relatifs à la Belgique sont ceux des naissances masculines légitimes, tandis que les données se rapportant à l'Italie visent les enfants nés vivants, d'une façon

(1) Les chiffres pour l'Italie sont reproduits d'après Benini, *Statistica metodologica*, p. 226, note 1. Les chiffres pour la Belgique viennent de l'*Annuaire statistique de la Belgique*.

générale. Ce fait ne change rien à la curieuse constatation d'une diminution constante (sauf une fois, en Belgique, entre 1891 et 1900) du rapport des naissances masculines. On voit donc que les phénomènes réputés les plus fixes sont susceptibles de subir des modifications lentes qui semblent les faire insensiblement dévier de la notion admise comme exacte. Le calcul des probabilités intervient ici encore pour préciser la valeur de ces déviations et faire reconnaître si leur ampleur est suffisante pour devoir faire rejeter l'hypothèse d'une probabilité constante exprimant un complexus de causes essentielles, dominant nettement les causes accidentelles. On a construit des tables qui servent à déterminer la probabilité que le phénomène reste contenu dans les limites de l'écart probable, c'est-à-dire un certain degré de déviation qui présente une égale probabilité d'être ou de n'être pas surpassée en un nombre donné d'épreuves et dont l'expression mathématique est :

$$\pm 0{,}4769 \sqrt{2\,m\,p\,q}$$

m étant le nombre des épreuves, p et q les probabilités simples des événements attendus, et 0,4769 représentant précisément le point médian des valeurs calculées allant de 1 p. c. à 99 p. c.

Le résultat du calcul étant connu, si l'on trouve que cet écart est constamment dépassé par rapport à la probabilité typique, on pourra légitimement soupçonner la présence d'une cause perturbatrice. Bien entendu, ceci suppose que l'on ait fait disparaître au préalable, au moyen de l'interpolation, les variations dépendant d'une cause générale.

En appliquant donc le calcul des probabilités à des séries suffisamment étendues, on arrivera à mettre en lumière leur caractère de stabilité ou d'instabilité, ce qui est encore un résultat essentiel dans les recherches statistiques (1).

---

(1) BENINI, *Statistica metodologica*, pp. 223 et 225.

48. Dans le domaine des faits influencés par la libre volonté de l'homme, le calcul des probabilités intervient utilement pour montrer jusqu'à quel point les résultats de l'expérience s'écartent de ceux du calcul théorique, c'est-à-dire pour déterminer la loi empirique du phénomène et permettre de comparer les tendances constatées avec les données absolues résultant du calcul. Dans le calcul des probabilités, on doit supposer que tous les événements, favorables ou défavorables, sont également possibles, c'est-à-dire qu'on les assimile au tirage de boules de différentes couleurs d'une urne qui contiendrait en nombre égal des boules blanches et des boules noires. Mais il existe des tendances qui, en pratique, manifestent leurs effets dans l'association ou la séparation de certains caractères, de telle sorte que la comparaison des deux résultats montre à quelle loi empirique obéit le phénomène considéré. Benini donne l'exemple suivant, tiré de la statistique italienne, concernant les combinaisons matrimoniales entre illettrés et illettrées. Nous résumons ses constatations présentées dans une forme aussi claire que sobre :

En 1896, les époux et les épouses capables de signer leur acte de mariage furent en Italie dans la proportion respective de 63,04 et de 47,44 p. c., ce qui laissait pour les hommes une proportion de 36,96 p. c. d'illettrés et pour les femmes une proportion de 52,66 p. c. Si on veut savoir quelle est la chance qu'un homme sachant signer son nom épouse une femme ne le sachant pas, ou réciproquement, ou que des illettrés se marient entre eux, ou que les combinaisons matrimoniales se forment entre personnes sachant l'une et l'autre écrire, on n'a qu'à calculer les résultats en appliquant le théorème de la probabilité composée, c'est-à-dire à multiplier entre elles les probabilités simples. Lorsqu'on compare les résultats du calcul théorique aux données de fait relevées par l'observation statistique, on constate

une forte différence, ainsi que le montre le petit tableau suivant :

COUPLES		Probabilité mathématique.	Observation statistique	Différences.
Groupes semblables	lettré et lettrée .	29,91 p. c.	42,44 p. c.	+ 12,53 p. c.
	illettré et illettrée	19,42 p. c.	31,96 p. c.	+ 12,54 p. c.
Groupes différents	lettré et illettrée .	33,13 p. c.	20,60 p. c.	— 12.53 p. c.
	illettré et lettrée .	17,54 p. c.	5,00 p. c.	— 12,54 p. c.

La probabilité mathématique est établie comme si les conditions sociales, qui sont plus ou moins exactement exprimées par ce degré élémentaire d'instruction permettant à l'intéressé de signer son acte de mariage, étaient des éléments indifférents. Mais il est démontré en démographie que les époux cherchent en général à s'associer dans des conditions semblables sous le rapport de l'instruction, de l'âge, de l'état civil, etc. Le tableau ci-dessus nous montre que l'attraction exercée par des conditions égales au point de vue de l'instruction s'exprime par une différence de 12 1/2 p. c. au-dessus ou au-dessous du calcul théorique.

Les exemples qui précèdent suffisent à préciser le genre de services que le calcul des probabilités peut rendre à la statistique en lui fournissant un point de comparaison fixe qui permet de dégager la loi empirique des phénomènes. Ce point apparaît comme d'une importance capitale dans l'interprétation. Il suffit de l'avoir souligné dans ces observations préliminaires qui ne doivent pas empiéter sur les développements ultérieurs de cet ouvrage.

49. Précisons simplement quelles sont les différences qui se marquent entre la probabilité *a priori* et la probabilité *a posteriori*, ou, comme dit M. Emile Borel, les différences entre la sphère et l'orange.

1° Dans la probabilité mathématique, on considère que les éléments qui entrent en ligne de compte sont parfaitement indépendants entre eux et qu'ils peuvent se combiner ensemble de toutes les façons possibles, de telle façon que le calcul du nombre de cas possibles s'opère d'après les règles indiquées plus haut.

Au contraire, la probabilité statistique ou *a posteriori* ne comporte pas le libre jeu de tous les éléments qui y interviennent. La présence de l'un peut amener la disparition de l'autre et provoquer l'arrivée d'un troisième.

C'est donc l'expérience qui doit nous éclairer en nous permettant de comparer les résultats de l'observation avec ceux du calcul théorique.

2° La base de la probabilité *a posteriori* se trouve dans un nombre *suffisant* d'observations réalisées selon les règles de la technique statistique. Mais quand un nombre d'observations est-il suffisant en soi, ou quand pouvons-nous le considérer comme offrant une base suffisante? On a, par exemple, l'habitude de diviser des phénomènes continus, comme ceux de l'ordre démographique, d'après des divisions du temps, correspondant à la vie civile, on compte les naissances, les mariages, les décès, par année. Qui nous dit qu'une année offre un terrain suffisant à l'observation pour baser sur les constatations effectuées durant ce temps une probabilité *a posteriori*? Cette division du temps que nous appelons une année a une importance réelle en astronomie et, pour mesurer les faits de la vie civile, elle est une unité logiquement choisie. Comment pouvons-nous savoir qu'elle convient également pour observer les faits sociaux dont nous avons à apprécier la probabilité de répétition? Remarquons, en passant, que la mesure du temps d'observation ne devrait pas être la même pour *tous* les faits : il y a un grand nombre de naissances et de décès chaque jour, donc au bout d'une année il y en a un nombre respectable, mais il y a moins de ma-

riages, moins encore de divorces. Pourtant, pour les uns comme pour les autres, on note ceux qui se sont présentés en une année.

3° La probabilité mathématique s'applique d'ordinaire à des faits simples; pour la probabilité statistique, les phénomènes complexes collectifs interviennent fréquemment. La possibilité d'une naissance masculine ou féminine est une probabilité simple, appuyée sur des observations extrêmement nombreuses, basées sur une loi générale; on peut espérer atteindre un degré de probabilité assez élevé à l'aide des éléments dont on dispose, pourvu qu'on n'attende pas la réalisation de cette probabilité dans un cas particulier, mais simplement comme caractéristique d'un grand nombre de cas envisagés ensemble. Il y a aussi des phénomènes complexes donnant lieu à des probabilités composées : le mariage d'un veuf avec une veuve, d'un célibataire avec une jeune fille, d'une jeune fille avec un veuf, d'une veuve avec un célibataire, sont des événements donnant lieu à des probabilités composées dont la réalisation est moins certaine, comme on l'a vu plus haut, que celle envisagée dans l'exemple précédent. La probabilité théorique, qui se déduit du calcul, est fréquemment en discordance avec les résultats de l'observation dans les phénomènes complexes, elle s'en rapproche davantage dans les phénomènes simples. Le degré de sécurité dans la prévision n'est donc pas le même pour tous les cas fondés sur la probabilité *a posteriori*.

4° Une dernière différence sépare la probabilité mathématique de la probabilité statistique : elle consiste dans la permanence des caractères qui sont la base de la première et dans les modifications constantes qui se remarquent dans les éléments constitutifs de la seconde. Lorsque je calcule combien de fois apparaîtra telle face des dés dans un nombre donné de parties, je suppose d'abord que toutes les faces de ces dés peuvent apparaître indifféremment et que

le jeu ne modifiera pas leur construction. Mais cette supposition, je ne puis la faire à propos des phénomènes sociaux relevés par la statistique, ou plutôt l'on peut faire la supposition opposée. Les faits démographiques, sociaux, économiques, se modifient constamment et les prévisions qui se basent sur l'observation ne peuvent être qu'à brève échéance. Si nous n'apercevons pas immédiatement les changements qui se sont produits, la raison en est à l'imperfection des moyens mis en œuvre pour observer et recueillir les faits. Avec la probabilité mathématique, nous sommes dans l'absolu; au contraire, la probabilité *a posteriori* est toute relative et n'autorise qu'une prévision qui devance timidement les faits; aussi l'observation statistique qui lui sert de base n'a-t-elle qu'une valeur limitée dans le temps et doit-elle être renouvelée à des intervalles assez proches.

III. — La statistique et les mathématiques.

50. C'est un sujet de vives controverses entre statisticiens théoriciens et professionnels que la question de savoir quelles sont au juste les relations entre la statistique et les mathématiques et ce qu'elles devraient être. Si opposées sont les vues, si tranchées sont les positions qu'il faut renoncer, lorsqu'on expose des idées de juste milieu, à se concilier les deux adversaires. Les professionnels trouveront sans doute qu'on aura trop cédé à un entraînement vers l'aspect scientifique pur, les théoriciens seront d'avis que la thèse soutenue est incomplète et regretteront de n'avoir pas été toujours suivis dans leurs revendications. Si ingrate que soit cette position de l'homme exposé aux feux croisés de ses adversaires, c'est pourtant celle que nous prendrons.

Certains mathématiciens sont catégoriques dans l'expression de leur opinion qui, parfois, prend les allures d'un

arrêt : « Les sciences statistiques, disait Fourier au début du XIX[e] siècle, ne feront de véritables progrès que lorsqu'elles seront confiées à ceux qui ont approfondi les théories mathématiques (1). »

Racioppi a soutenu que ce sont les mathématiques seules qui ont élevé la statistique au rang de science (2). Un mathématicien français, M. H. Laurent, soutient la même thèse avec beaucoup d'animation; selon lui, « une erreur très répandue consiste à croire que celui qui dirige des opérations de statistique n'a pas besoin de connaître les mathématiques, le calcul des probabilités et en général les sciences auxquelles correspondent les chiffres qu'il interprète (3) ». Aussi, comme cette erreur est surtout celle des gouvernants, « nous ne possédons pas de bonnes statistiques officielles parce que les personnes qui dirigent ces statistiques ne sont pas bien préparées à ce genre de travail (4) ». Où pourrait-on recruter des personnes bien préparées? Parmi les membres de l'Institut des actuaires français, répond M. Laurent. « C'est dans le sein de cette société que le gouvernement trouverait des sujets aptes à diriger des statistiques officielles (5). » Voilà qui est bien catégorique. Les statisticiens qui étaient en même temps mathématiciens ont montré plus de discrétion. Messedaglia se borne à souhaiter que parmi les statisticiens il ne s'en trouve plus un seul qui soit ignorant en mathématiques (1). Ferraris est plus réservé encore : « La statistique réclame, dit-il, de

(1) *Correspondance mathématique*, t. II, p. 177. Bruxelles, 1826.

(2) RACIOPPI, *Del principio e dei limiti della statistica*, p. 108. Napoli, 1857.

(3) H. LAURENT, *Statistique mathématique. Avertissement*, p. II. Paris. 1910.

(4) Id., *loc. cit.*, p. II.

(5) Id., *loc. cit.*, p. IV.

(6) MESSEDAGLIA, *Esposizione critica delle statistiche criminali dell'Imperio Austriaco*, p. 151.

qui veut la posséder à fond, une certaine connaissance (*una discreta cultura*) des sciences mathématiques (1). »

51. Il est intéressant d'entendre une autre opinion exposée par des hommes éminents, ayant occupé les plus hautes fonctions dans la statistique administrative et ayant à leur actif une longue série d'œuvres statistiques, dont quelques-unes d'allure monumentale. Voici l'avis de M. Bodio, exprimé dans une circonstance solennelle, le discours de clôture de la vingt-troisième session de l'Institut international de statistique : « Les questions de méthode, dit le vénéré président de l'Institut, ont amené un échange de vues entre les statisticiens au sujet de l'emploi des mathématiques supérieures. Certes, les mathématiques sont un moyen puissant d'analyse pour affiner le matériel de la statistique et un instrument de synthèse pour en condenser les résultats et en tirer de nouvelles inductions. Le danger est que les mathématiciens de profession, pour l'amour de l'art, sont amenés parfois à trop subtiliser et à déduire des conséquences dépassant, peut-être, ce que la qualité des observations pourrait permettre; ils appliquent à ces observations le trébuchet du peseur d'or, alors que ce qui leur convient, c'est la balance grossière du marchand. »

« Ce qui facilite la solution des difficultés, c'est que, pour certaines branches, notamment pour la biométrie, il s'est formé presque une profession spéciale, celle des actuaires. Pour les enquêtes d'ordre économique, le calcul supérieur n'est pas précisément nécessaire; on demande surtout un jugement sûr, un esprit critique et une impartialité absolue, dégagée de toute préoccupation d'école (2). »

---

(1) FERRARIS, *La statistica et la scienza dell' Amministrazione nella facolta giuridiche*.

(2) *Bulletin de l'Institut international de statistique*, XIIIe session, T. XIX, première livraison, p. 154.

La thèse soutenue par M. von Mayr dans son grand traité (1) ne diffère pas sensiblement de celle-là. Comme M. Bodio, il reconnaît aux procédés mathématiques l'avantage d'une plus grande élégance, mais loin d'admettre que l'emploi d'une technique supérieure soit indispensable, il déclare que « tous les groupements des éléments statistiques et tous les calculs qui s'y rapportent peuvent s'exprimer dans le langage usuel de l'arithmétique ». M. von Mayr ajoute, pour fortifier sa thèse, que le calcul des probabilités n'est admissible, en statistique, qu'à l'égard des phénomènes de la vie sociale, qui présentent une certaine analogie avec la répétition de l'observation d'un même objet ou la chance de retirer une certaine bille d'une quantité limitée et constante de billes de différentes couleurs (2).

« Parmi les phénomènes sociaux les plus importants, dit-il encore, il en est pour lesquels c'est le changement perpétuel des masses qui est caractéristique ; ici, certainement, le calcul des probabilités n'a pas à intervenir ». Et comme conclusion, on peut retenir ces lignes empruntées au même auteur : « Une méthode d'étude qui n'est accessible qu'à un cercle restreint d'initiés et qui, de plus, n'est pas absolument indispensable, ne peut constituer une branche intégrante de la statistique, considérée comme une science des-

---

(1) G. VON MAYR, *Statistik und Gesellschaftslehre*, Fribourg, 1895, p. 615. Cf. également la nouvelle édition de cet ouvrage, Tubingen, 1914.

(2) « D'après les résultats acquis actuellement, dit M. von Mayr, l'application du calcul des probabilités n'accuse des indications spéciales quelque peu satisfaisantes — encore ne s'agit-il que de données plus ou moins variables — que dans le cas où les unités considérées ont plutôt le caractère d'un fait physique que d'un fait social ; ainsi, la grandeur des individus, les naissances, la répartition des âges, sont de simples faits naturels qui se prêtent au calcul. Mais partout où l'on se trouve en présence de faits purement sociaux, l'application du calcul des probabilités ne donne rien. Les phénomènes de la vie sociale ne sont pas stables comme ceux de la nature : ils sont sujets à un développement infini et incalculable, lequel précisément est mis en évidence par l'étude des collectivités pratiquée en statistique. La valeur de l'objet de cette science n'est diminuée en rien par le fait qu'il ne se laisse pas figer en des formules mathématiques. »

tinée à la généralité des esprits cultivés. On peut préférer la méthode mathématique, mais celui qui n'en possède pas les éléments ne doit pas pour cela renoncer à étudier les problèmes qui constituent le domaine de la statistique. »

52. Ces opinions, si profondément divergentes, ne paraissent pas laisser de place à un jugement intermédiaire. Nous essayerons cependant de trouver ce terrain mixte entre les affirmations des uns et les dénégations des autres.

Un premier point qui nous semble, à cet égard, d'une grande importance est qu'il convient de distinguer entre les deux grandes phases des opérations statistiques : le relevé et le dépouillement d'une part, l'exposition et l'interprétation d'autre part. C'est une idée qui se trouve exposée par un mathématicien-statisticien distingué, le professeur A. Bowley, dans une étude très suggestive des rapports entre la statistique et les mathématiques (1).

Pour M. Bowley, le travail arithmétique est limité à la tabulation des chiffres concrets, là où le champ d'observation peut être couvert tout entier, là où l'approximation et l'interpolation ne sont pas nécessaires et où la statistique ne diffère guère que par le nom de la comptabilité. Au contraire, les procédés mathématiques permettent de mesurer ce qui, sans cela, resterait inaccessible, de décrire, comme en paléontologie, l'animal à l'aide d'un seul de ses os, d'établir des observations fermes, sur des bases mouvantes. C'est un véritable microscope qui permet de mesurer des différences que la méthode arithmétique ne peut apercevoir, c'est un procédé qui permet de raisonner correctement d'après les résultats de la tabulation et de discerner le fait essentiel obnubilé dans les différentes manifestations des phénomènes.

---

(1) A.-L. Bowley, « Address to the Economic Science and Statistic section of the British Association for the advancement of Science ». York, 1906. (*Journal of the Statistical society*, 1906.)

La distinction établie par M. Bowley est exacte, à la condition de compléter l'exposé de l'auteur sur un point qui ne manque pas d'importance. M. Bowley a eu en vue, d'une part, les opérations matérielles qu'il compare, non sans raison, à une comptabilité (mais de quelle complication !) et, d'autre part, les travaux plus savants de la synthèse des résultats et de l'interprétation. Cependant, on peut se demander comment le mathématicien trouverait matière à ses développements, si son point de vue mathématique lui-même n'avait pas été à la base du travail, si l'on n'avait pas pris soin de rassembler des matériaux de telle nature et dans de telles conditions, que l'on pût formuler plus tard les équations ? En d'autres termes, il ne peut jamais exister de dissociation entre le travail préliminaire et le travail final. Celui qui, sur un sujet donné, prend la plume pour tracer le plan ou programme des opérations, doit se demander tout d'abord à quelles conditions doit répondre le travail. Si parmi elles il s'en trouve — c'est le cas ordinaire — qui aient rapport aux mathématiques, l'auteur du plan doit savoir quels sont les éléments nécessaires pour établir les formules afin que les recherches à instituer fournissent les données réclamées par l'actuaire. Ceci est l'évidence même et nous ne doutons pas que telle soit, au fond, la pensée de M. Bowley, mais elle n'aura rien perdu à être exposée d'une façon un peu plus explicite.

En dehors de la condition générale que nous venons de souligner, établir les bases d'un questionnaire, rédiger les instructions à l'usage des agents recenseurs, arrêter une classification, examiner les réponses au point de vue de leur sincérité et de leur exactitude, dresser les modèles de tableaux statistiques, dépouiller les résultats de l'enquête, en consigner les résultats dans les tableaux, procéder aux opérations arithmétiques qui sont la suite et le prolongement naturel du dépouillement, tout cela relève de l'économie politique et sociale, de la méthodologie statistique

générale et spéciale et de l'arithmétique. Mais on ne peut rien y trouver qui appartienne aux mathématiques. Il est vrai que M. Bowley suppose qu'il ne s'agit pas de se livrer à des calculs d'interpolation (1), mais, en pratique, ces cas ne se présentent que rarement. C'est précisément une des conquêtes de la statistique moderne d'avoir réduit dans des limites très restreintes les cas dans lesquels le relevé ne peut porter sur tous les objets faisant partie du domaine de l'observation. Le statisticien professionnel, appliqué aux recherches usuelles, ne se trouve pas, le plus souvent, dans l'occasion d'appliquer des méthodes relevant du calcul supérieur. Mais ce n'est pas un motif de sous-évaluer le travail auquel il se livre; l'examen que nous ferons plus loin des difficultés de la statistique pratique montre assez qu'il ne s'agit pas d'une besogne peu qualifiée, digne d'être abandonnée à des *unskilled*, tandis que l'homme de science se bornerait à utiliser les matériaux réunis par des manœuvres. Le divorce entre l'élément scientifique et l'élément pratique aurait au contraire les plus funestes conséquences.

L'importance de la technique est énorme. N'est-ce pas elle qui apporte à la science les matériaux d'étude que celle-ci réclame? Toute l'histoire de la statistique n'est-elle pas là pour montrer qu'on s'est épuisé en efforts pour améliorer la technique? Et l'œuvre d'une institution comme l'Institut international de statistique, qu'est-elle sinon un effort incessant pour améliorer les conditions du relevé et de la tabulation?

---

(1) Le calcul par interpolation consiste à rechercher, en l'absence de données obtenues par l'observation directe, une ou plusieurs mesures intermédiaires situées entre d'autres résultant du relevé. Il sert aussi à rectifier certaines mesures qui, bien qu'obtenues par les recherches statistiques, sont certainement entachées d'erreur. Enfin, il peut être nécessaire, afin de comparer entre elles des statistiques qui ne sont pas présentées d'après une même base ou échelle.

L'interpolation peut être faite par un procédé graphique ou par l'algèbre (calcul différentiel).

Quel serait le rôle des mathématiques dans l'interprétation? C'est encore à M. Bowley que nous demanderons la réponse. On peut, dit-il, concevoir l'application des mathématiques à la statistique comme un instrument de précision dont l'usage a pour but de faire apparaître des points qui, sans cela, resteraient obscurs. C'est dans ce sens qu'il compare la méthode mathématique au microscope, tandis que l'arithmétique ne dépasserait pas la portée de l'œil de l'observateur. On peut dire aussi que l'emploi des mathématiques supplée au défaut d'observation, au manque de continuité dans une observation commencée, interrompue et reprise ensuite. Enfin, ainsi qu'il ressort à toute évidence de la critique des procédés statistiques, le matériel réuni est habituellement imparfait et sa mise en œuvre ne va pas sans courir quelques chances d'erreur; si excellents que soient les procédés techniques, il faut toujours compter avec les erreurs commises par les aides nombreux, à la collaboration desquels il faut recourir pour la mise en œuvre de tout matériel un peu important. Or, le matériel étant imparfait, c'est une des exigences du travail scientifique de faire disparaître ses tares dans la plus large mesure possible ou d'évaluer avec précision celles que nous ne parvenons pas à éliminer. C'est surtout dans ces deux directions, dit M. Bowley, que les méthodes des mathématiques sont généralement nécessaires. Le matériel est mis à l'épreuve par les méthodes d'interpolation et de graduation. La loi générale de groupement et de direction du mouvement est découverte et les variations accidentelles sont éliminées, ou inversement, la direction générale est neutralisée et les variations accidentelles mesurées.

Le rôle des mathématiques dans la statistique consisterait donc : *a*) à combler les lacunes de relevé: *b*) à améliorer les résultats de l'observation et à évaluer la grandeur des erreurs qui ne peuvent être corrigées; *c*) à affiner les conclusions de l'interprétation en faisant apparaître des

éléments qui, autrement, resteraient inconnus ou imparfaitement connus.

53. Il est intéressant de reprendre en détail à cet endroit les objections que Mayr adresse aux mathématiciens.

Pour le premier point, Mayr se borne à remarquer, succinctement, que le champ d'application ouvert aux mathématiques se restreint de plus en plus, grâce au caractère toujours plus complet et plus général des observations primaires; quant à l'extrapolation, elle n'a d'autre importance que celle qu'on peut attacher à d'ingénieuses combinaisons, sans grande portée pratique (1).

On emploie les mathématiques pour découvrir les erreurs accusées par l'observation, mais le mot « erreur » a ici un sens particulier. Il est un point, dit cet auteur, qu'en tout cas l'étude la plus attentive par les mathématiques pures est incapable d'élucider : c'est de savoir s'il s'agit d'erreurs matérielles de l'observation, ou si ces écarts, qui paraissent des erreurs au sens des mathématiques, ne sont pas simplement des écarts que présente la réalité par rapport à une normale théorique (2).

Ce langage semble inspiré par quelque parti-pris, il ne paraît pas concorder très exactement avec la réalité des choses. Les procédés d'interpolation rendent d'incontestables services dans tous les cas, fréquents en statistique, où il s'agit d'éliminer l'effet de causes constantes pour laisser apparaître seulement celui des causes secondaires et accidentelles. Des rapports de grandeur et de variabilité ne peuvent parfois être mis en lumière d'une façon complète qu'à l'aide de procédés mathématiques. Enfin, il n'est pas niable que dans l'interprétation, le statisticien puisse recueillir des avantages sérieux de la connaissance des mathématiques.

---

(1) von Mayr, *loc. cit.*, paragr. 15, n° 2.

(2) *Ibid.*, paragr. 15, n° 5.

En conclusion, la thèse négative pure et simple, telle qu'elle est professée par plusieurs auteurs, ne paraît pas soutenable dans les circonstances actuelles. Etant données les applications des mathématiques à la statistique, il n'est pas possible de s'en désintéresser. La formule de Ferraris, qui souhaite que chaque statisticien ait une connaissance suffisante (*una discreta cultura*) des sciences mathématiques, peut être considérée comme un minimum, même à l'égard de l'organisateur professionnel des recensements et des enquêtes.

54. En effet, il n'est pas douteux que les procédés mathématiques sont de nature à élucider bon nombre de problèmes; ils font dès lors partie intégrante de la technique et le statisticien doit être à même d'y recourir, comme il recourt aux procédés techniques d'autre nature. Lorsque la technique d'une science évolue, nul n'a le droit de se désintéresser des acquisitions nouvelles dûment éprouvées. Que dirait-on d'un biologiste qui, obstinément fidèle aux méthodes qui lui furent enseignées durant sa jeunesse, refuserait d'appliquer, même de connaître, les procédés techniques nouveaux découverts depuis? Le statisticien qui fermerait les yeux à l'évidence, se mettrait dans une situation peu défendable. Bien entendu, il y a des degrés résultant de la nature des choses; on pourra demander davantage au démographe qu'au statisticien économiste, par exemple. Ce qui est hors de doute, c'est que chaque statisticien doit posséder l'*esprit* mathématique, cette disposition, peut-être indéfinissable, qui fait que l'on se rend compte de l'exactitude mathématique d'une proposition et que l'on suit avec intérêt les développements donnés par les techniciens. Ni l'éloignement ni l'indifférence ne sont plus de mise; les mathématiques ont fait leurs preuves, elles font partie de la technique; on ne peut plus s'en désintéresser, car il serait contraire à l'esprit scien-

tifique de laisser à l'abandon une parcelle quelconque de la méthode capable de conduire à la vérité.

55. Une autre question surgit aussitôt. Dans quelle mesure cette méthode est-elle « indispensable » ? On peut répondre, nous paraît-il, qu'elle n'est pas toujours « indispensable », tout en étant très souvent utile. D'après les termes mêmes du problème, tel qu'il a été posé par M. Bowley dans son adresse présidentielle à la British Association, en 1906, l'un des buts essentiels de la méthode mathématique consisterait à combler les lacunes du relevé. Mais une objection peut être faite : les mathématiques ne sont pas de la divination; lorsqu'elles essayent de suppléer au relevé, — soit que celui-ci soit incomplet ou qu'il soit discontinu, ou que l'on tente de transporter dans l'avenir les conclusions pour le présent, — on part toujours de ce postulat que les conditions connues s'appliquent aux phénomènes restés en dehors du relevé; or, c'est précisément ce qu'on ignore. Le plus souvent, on se trouve en présence de faits essentiellement variables et différents les uns des autres : de la grandeur des uns, bien connue, on ne peut conclure à la grandeur des autres, qui est inconnue. Ne trouve-t-on pas plus de garanties dans la tendance, si accentuée aujourd'hui, d'étendre le relevé à toutes les unités? Vérifier si les résultats du relevé sont complets, s'ils ne présentent pas de trop grandes divergences avec les résultats antérieurement recueillis, comparer les déclarations des ouvriers avec celles des patrons, veiller à éliminer les multiples emplois si faciles à commettre quand plusieurs personnes peuvent se croire qualifiées pour répondre au questionnaire, — toutes ces précautions concourent au même but : s'assurer que le relevé est complet, empêcher que la même unité ne soit comptée plusieurs fois. Elles exigent de la patience, du travail, de l'expérience, mais elles n'ont rien à voir avec les mathématiques. Ceci ne signifie pas que la méthode ma-

thématique est, dans ce cas, inutile ou mauvaise, mais simplement qu'elle n'est pas toujours indispensable.

56. L'emploi des procédés mathématiques peut avoir un troisième but : celui d'affiner les conclusions de l'interprétation en faisant apparaître des éléments qui, autrement, resteraient en tout ou en partie inconnus. C'est cette fonction des méthodes mathématiques que M. Bowley compare à l'emploi du microscope et c'est à propos d'elles qu'il dit que la méthode arithmétique n'a qu'une portée relative comme une observation faite à l'œil nu ne nous révèle qu'un aspect incomplet des choses.

La légitimité de l'emploi des méthodes mathématiques dans le sens indiqué n'est pas contestable, mais son usage judicieux présente de sérieuses difficultés. C'est ici surtout qu'est exacte la parole de Bodio qui craint de voir employer, pour l'amour de l'art, le trébuchet du peseur d'or quand on ne devrait recourir qu'à la balance grossière du marchand. Les applications parfois subtiles qu'on fait du calcul supérieur cadrent mal avec l'aspect fruste des données numériques. Avec les résultats des statistiques, nous nous trouvons dans le domaine du relatif, avec le calcul mathématique, nous sommes dans la sphère de l'absolu; existe-t-il entre ces deux aspects des choses un pont solide qui permette de passer de l'un à l'autre? Il sera toujours sage de poser la question.

Il n'en reste pas moins vrai que les procédés arithmétiques ne permettent pas toujours de discerner toutes les faces d'une question. Ainsi que nous le montrerons plus loin, les relations qui existent entre les manifestations quantitatives de deux phénomènes ne sont complètement mises en lumière qu'avec l'aide du calcul des corrélations. M. Udny Yule considère la méthode de corrélation comme fondée sur la formation d'équations donnant la valeur d'une variable en fonction d'une ou de plusieurs

autres. La méthode de corrélation consisterait donc à déterminer le degré d'affinité qui unit entre eux deux ou plusieurs éléments dont les relations sont susceptibles de mesure. Lorsque ces éléments sont unis entre eux par des rapports de grandeurs, le degré de persistance de la relation qui existe entre ces grandeurs caractérise entre elles le degré de corrélation. Les éléments sont alors appelés fonctions les uns des autres, c'est-à-dire qu'il existe respectivement entre eux une dépendance plus ou moins parfaite qui les fait varier de grandeur en raison des variations subies par les autres éléments dont ils sont fonctions.

La méthode de corrélation est essentiellement mathématique, elle ne peut être remplacée par aucune opération de l'arithmétique. A ce point de vue, il est exact que les procédés mathématiques font apparaître des résultats qui, sans eux, ne pourraient être mis en lumière.

CHAPITRE IV

Division de la matière.

57. Il semble que nous sommes arrivés au point où la matière peut se diviser d'après des caractères logiques.

En premier lieu, nous distinguons la *statistique méthodologique*, qui comprend l'ensemble des procédés par lesquels on parvient à la connaissance des phénomènes collectifs sans distinction de leur nature. Qu'il s'agisse de faits sociaux, biologiques ou physiques, les observations se font au moyen du relevé, sont soumises à la critique, sont exposées dans des tableaux, puis condensées en formules abréviatives et, enfin, interprétées selon les mêmes procédés logiques. Sans doute, les modalités diffèrent selon la nature des objets, mais pour le fond, la méthode reste identique. On ne saurait trouver de distinction logique entre un relevé

des jours de pluie, le dénombrement du bétail et un recensement de l'activité économique de la population. Ce qui sépare ces travaux — en ne considérant que le problème méthodologique — c'est surtout le degré de difficulté que leur exécution peut présenter. Dans le premier cas, le pluviomètre fournira peut-être des indications suffisamment précises, on se bornera à noter régulièrement les résultats enregistrés par l'instrument, relevé commode et sûr, auquel on peut assimiler certaines formes du relevé automatique (1). S'agit-il de compter le bétail? La chose est plus compliquée, car il faut en venir à un comptage des unités; nulle opération de l'espèce ne va sans quelque risque d'erreur : omissions, ou, par contre, double emploi, inattention des employés, erreurs de calcul, etc. Mais, comme ce n'est qu'un comptage par espèce, il s'agit d'une opération d'une difficulté médiocre. En vient-on au contraire au recensement professionnel ? Voilà que surgissent bien d'autres problèmes! Les qualités économiques sont ardues à définir, plus difficiles encore à discerner; il ne faut que peu de chose — un mot mal employé, une expression à double entente — pour faire naître de graves méprises; les risques d'omission ou de répétition sont nombreux; de plus, ils passent facilement inaperçus. Voilà pour les différences. Elles sont notables. Mais, en somme, la base reste toujours la même : c'est le relevé, c'est-à-dire « l'observation, unité par unité, des manifestations individuelles d'un phénomène et des conditions de toutes ses manifestations ».

La statistique méthodologique est donc générale par son objet. Elle résume la technique particulière de la méthode, trace la voie à suivre, indique les précautions à prendre pour arriver à de bons résultats, peu importe la nature des faits auxquels on l'applique.

---

(1) Cfr. livre I, 1re section, ch. I, § III.

58. A côté de la statistique méthodologique, nous plaçons la *statistique descriptive*. On a essayé à différentes reprises d'en tracer un plan systématique. Ferraris, par exemple, considérant la société humaine comme un organisme qui se reproduit, se développe et meurt, a proposé la division suivante (1) :

I. Notions préliminaires comprenant l'étude anthropométrique et psychique de l'individu.

II. Production de la vie (nuptialité, natalité, divorce, naissances illégitimes, mort-nés).

III. Distribution de la vie (densité de la population).

IV. Circulation de la vie (mouvement de la population, enseignement, religion).

V. Consommation de la vie (maladies, suicides, criminalité).

VI. Epilogue (augmentation ou diminution de la population, libre arbitre individuel).

Messedaglia se rapproche beaucoup de Quetelet (2) dans les divisions suivantes qu'il propose d'adopter pour l'étude et le partage des matières à considérer par la statistique :

I. Territoire ou topographie.

II. Population ou démographie.

III. Vie économique : agriculture, industrie, commerce, voies et moyens de communication.

IV. Vie intellectuelle : instruction publique à tous les degrés, presse, bibliothèques, etc.

V. Vie morale : criminalité, prostitution.

VI. Vie politique : statistique judiciaire, financière, militaire, électorale, etc.

59. Nous pensons qu'il convient de réserver une place spéciale à la démographie. Pour beaucoup de personnes, la démographie est inséparable d'un exposé réservé à la statistique méthodologique. On la considère souvent comme le prolongement nécessaire, obligé de celle-ci, comme si la méthodologie formait la partie générale, la démographie la partie appliquée ou spéciale d'une science unique : la statistique. C'est une erreur. Levasseur a parfaitement mis

---

(1) C. Ferraris, *La statistica; le sue partizioni teoretiche*. Venezia, 1890

(2) Cf. plus haut, n° 16.

en lumière le caractère propre de la démographie : « Il y a, dit-il, une étude qui constitue assurément une science et qui est si étroitement liée avec la statistique, qu'on l'a confondue souvent avec elle : c'est la démographie. On a désigné et on désigne encore la démographie sous le nom de « statistique de la population », expression à peu près exacte, quoique un peu étroite, ou simplement de « statistique », expression inexacte. La démographie est la science de la population ; elle en constate l'état, elle en étudie les mouvements, principalement dans la naissance, le mariage, la mort et dans les migrations, et elle s'efforce de parvenir jusqu'à la connaissance des lois qui la régissent. C'est la science de la vie humaine dans l'état social ; c'est bien réellement une science dans le sens que nous donnons à ce mot, puisqu'elle a un objet distinct, nettement déterminé. « Si elle a été parfois confondue avec la statistique, c'est que cette dernière est la méthode par laquelle elle procède dans la plupart de ses investigations et la mine de laquelle elle tire la plupart de ses matériaux (1). »

Ces deux dernières observations de Levasseur, qui ont rencontré l'adhésion d'un grand nombre d'auteurs, sont frappées au coin du bon sens. La démographie ayant son objet propre, sa sphère parfaitement délimitée, est une science qui se sépare nettement de la statistique méthodologique. Si la statistique lui fournit la plupart de ses données, elle ne les lui procure pas toutes. La démographie fait de larges emprunts aux sciences anthropologiques et biologiques, aux sciences politiques et administratives, aux sciences juridiques et économiques. En ne se soumettant pas à une méthode unique, elle achève de se classer parmi les sciences et se sépare des disciplines qui n'utilisent qu'un seul procédé de recherche.

---

(1) E. Levasseur, *La population française*, t. I, p. 18. Paris, Rousseau, 1889.

On pourrait en dire autant des recherches désignées sous le nom de *statistique morale*. Ses relations avec la démographie sont étroites, mais au lieu de l'état et du mouvement de la population, la statistique morale envisage certaines actions humaines sous le rapport de l'éthique. Les divorces, les suicides, la criminalité, la folie, la prostitution, sont les principaux sujets de ses recherches. On a proposé de la définir : « l'étude statistique de l'immoralité, du désordre moral(1) ». Elle a donc bien aussi son objet propre, son objectif spécial et limité ; si, comme la démographie, elle fait surtout appel à la méthode statistique, celle-ci n'est pas seule à lui fournir son concours. Pour apprécier le phénomène collectif qu'on désigne sous le nom de statistique morale, il ne suffit pas de la statistique, mais il faut encore utiliser, dans une large mesure, l'anthropologie, la psychologie et le droit. Un objet particulier, un faisceau de méthodes concordantes vers un but unique assurent à la statistique morale, au même degré qu'à la démographie, le caractère de la science.

60. En dehors de ces deux sciences bien définies, la démographie et la statistique morale, il reste une masse énorme de matière qui constitue la *statistique descriptive* à proprement parler. On peut la diviser en statistique du territoire, statistique économique, statistique de la vie intellectuelle, politique, administrative, statistique appliquée aux phénomènes physiques, aux caractères biologiques, etc.

Ce qui caractérise la statistique descriptive, c'est d'abord son étendue qui interdit au chercheur isolé d'embrasser la matière dans son ensemble. C'est ensuite la prédominance du point de vue méthodologique sur les résultats concrets. Ceux-ci sont sujets à d'infinies modifications que

(1) Camille Jacquart. « Essais de statistique morale. La criminalité belge 1868-1909 », p. 5. (Extrait des *Annales de l'Institut supérieur de philosophie*), Louvain. 1912.

la statistique doit relever avec patience et commenter avec sagacité. Mais dans un travail consacré aux procédés statistiques, il semble bien que le problème méthodologique se pose avant tout autre : montrer comment on fait de bonnes observations dans une sphère nettement définie est chose plus importante que de rassembler dans des tableaux, à coup sûr ingénieux, mais d'un intérêt tout momentané, des résultats dont la comparabilité est parfois douteuse. Il semble donc que la statistique descriptive ait à se proposer de noter l'évolution des méthodes et les acquisitions de la science au lieu de ne viser qu'à réunir des tableaux de chiffres et à en tirer des commentaires.

Nous exposerons en premier lieu les principes de la *statistique méthodologique,* puis nous en verrons les principales applications à la *statistique économique* et à la *statistique du travail.* Cet ouvrage se divise donc en deux parties : la première, consacrée à la méthodologie, est générale; la seconde, qui concerne la statistique économique, est spéciale.

Nous plaçant surtout au point de vue méthodologique, nous ne traiterons point des deux sciences dont nous avons reconnu l'existence indépendante : la *démographie* et la *statistique morale.* Ces sciences ont du reste fait l'objet de nombreux et savants travaux.

---

61. — *Références.*

BENINI (R.), *Principii di Demografia.* Firenze, B. Barbera, 1901.

ID., *Principii di statistica metodologica.* Torino, Unione tipografica editrice, 1906.

BERTILLON (J.), *Cours élémentaire de statistique administrative.* Paris, 1896.

BLOCK (M.), *Traité théorique et pratique de statistique.* Deuxième édition. Paris, Guillaumin, 1886.

BOREL (E.), *Eléments de la théorie des probabilités* (Cours de la Faculté des sciences de Paris). Paris, Hermann, 1909.

BOSCO (A.), *Lezioni di statistica. Parte prima; metodologica statistica.* Roma, Ermanno Loescher e C°, 1909.

BOWLEY (A), *Address to the Economic science and statistic section of the British Association for the advancement of science.* York, 1906.

ID., *Eléments of statistics.* Second edition. London, P. S. King and Son, 1902.

COLOJANNI (N.), *Lezioni di statistica.* Napoli, 1903.

COURNOT (A.), *Exposition de la théorie des chances et des probabilités.* Paris, Hachette, 1843.

DAVENPORT, *Statistical methods with special reference to biological variation.* London, Chapman, 1904.

DUFAU (P.-A.), *Traité de statistique.* Paris, 1840.

FAHLBECK (P.-E.), *La régularité dans les choses humaines ou les types statistiques et leurs variations* (*Journal de la Société de statistique de Paris,* 4e année, 1900), p. 188.

FAURE (F.), *Eléments de statistique,* Paris, Larose, 1906.

FERRARIS (C.), *La statistica e la scienza dell' Amministrazione nelle facoltá giuridiche.* Padova, 1878.

GABAGLIO (A.), *Teoria generale della statistica.* Milano, Ulrico Hœpli, 1888.

JACQUART (C.), *Statistique et science sociale* (aperçus généraux). Desclée, Bruxelles, 1907.

ID., *Essais de statistique morale. La criminalité belge,* 1868-1909. Louvain, 1912.

JEVONS (W. Stanley). *The principles of science* (*a Treatise on logic and scientific method.*) London, Macmillan & C°, 1909.

JULIN (A.), *Précis du cours de statistique générale et appliquée,* quatrième édition. Bruxelles, De Wit; Paris, Marcel Rivière, 1919.

KING (W.), *The elements of statistical method.* New-York, Macmillan, 1912.

LAURENT (H.), *Statistique mathématique.* Paris, 1910.

LEVASSEUR (E.), *La population française,* t. 1. Introduction sur la statistique. Paris, Rousseau, 1889.

LIESSE (A.), *La statistique, ses difficultés, ses procédés, ses résultats*. Paris, 1905.

MARCH (L.), *Remarques sur la terminologie en statistique* (*Journal de la Société de statistique de Paris*, 49e année, 1901, p. 290).

ID., *De la méthode en statistique* (de la méthode dans les sciences, 2e série). Nouvelle collection scientifique, directeur E. Borel. Paris, 1911.

MAYO-SMITH, Article « Statistical method » dans le dictionnaire de Palgrave.

MAYR e SALVIONI, *La statistica e la vita sociale*, seconda edizione. Torino, Ermanno G. Loescher, 1886.

MAYR (G. von), *Statistik und Gesellschaftslehre. Erster band, Theoretische statistik.* Zweite Auflage. Tübingen, 1914.

MEITZEN (A.), *History, theory and technique of statistics.* Traduction anglaise de Roland P. Falkner. Philadelphia, 1891.

QUETELET (A.), *Lettres sur la théorie des probabilités appliquées aux sciences morales et politiques.* Bruxelles, Hayez, 1846.

RACIOPPI, *Del principio e dei limite della statistica.* Napoli, 1857.

RUMELIN (G.), *Problèmes d'économie politique et de statistique.* Traduction française de A. de Riedmatten. Paris, Guillaumin, 1896.

TAMMEO (G.), *La statistica.* Torino, Roux, Frassatti et Cie, 1896.

VENN (J.), *The logic of chance*, 3e édition. London, Macmillan, 1888.

VERRIJN-STUART (C.-A), *Inleiding tot de beoefening der statistiek;* premier vol. Bohn, Haarlem, 1910.

VIRGILII (F.), *Statistica*, cinquième édition. Milano, Hoepli, 1911.

WAGNER (A.), *Del concetto, dei limiti e dei mezzi di esecuzione della statistica* (traduzione di Rodolfo Erny). Mémoire original publié en 1867 dans le *Staatswörterbuch*, de Bluntschli. *Annali di statistica*, série 2, vol. 7.

WORMS (R.), *La statistique*, Revue internationale de sociologie, 1904, no 7.

LIVRE PREMIER
TECHNIQUE DU RELEVÉ STATISTIQUE.

PREMIERE SECTION
Le relevé statistique ou relevé direct.

CHAPITRE PREMIER
Généralités, Définition, Division.

I. — Définition du relevé statistique.

62. La statistique est une méthode scientifique basée sur l'observation; elle relève entièrement de l'induction : elle part de faits constatés d'après des règles spéciales; elle raisonne sur des groupes homogènes formés à l'aide des phénomènes qu'elle considère; elle s'efforce ensuite de parvenir à la connaissance générale des tendances et des régularités qui se manifestent parmi les faits collectifs. La première phase de toute opération statistique consiste donc dans la préparation, l'exécution et la centralisation des observations faites sur les phénomènes collectifs.

La recherche statistique porte un nom spécial : on l'appelle le *relevé statistique* et ce terme convient mieux que celui d'*observation,* qui a été parfois employé. L'observation est la forme intellectuelle de l'attention; même elle est sa forme scientifique, car l'observation n'a sa raison d'être que dans le désir de connaître, de vérifier des idées générales. Elle se distingue nettement de l'expérience, par le caractère propre de sa méthodologie : l'observation est l'examen, l'investigation des phénomènes tels qu'ils sont,

— l'expérience est l'investigation d'un phénomène modifié par l'expérimentateur (1). Or, si nous examinons le processus statistique, pouvons-nous dire qu'il est un procédé d'observation? A proprement parler, le statisticien n'observe pas; il se borne à utiliser des observations faites par d'autres, sous sa direction, ce qui exclut le caractère d'attention personnelle se trouvant à la base de l'observation scientifique.

Cette raison ne paraît cependant pas entièrement décisive. Il suffit d'étendre un peu le sens du mot « statisticien », et d'y comprendre les agents du relevé, pour répondre à l'objection; d'ailleurs, les agents employés aux observations statistiques n'agissent que d'après l'impulsion et selon les méthodes de l'organisateur de la recherche, leur travail est simplement une projection du sien. Mais il y a, semble-t-il, une meilleure raison de préférer le mot de *relevé* à celui d'*observation*. Comme forme de l'investigation scientifique, l'observation doit pouvoir considérer tous les aspects extérieurs des choses auxquelles elle s'applique. Or, il n'en est pas ainsi dans l'observation statistique : elle peut relever seulement certains caractères des faits, ceux susceptibles d'être exprimés sous une forme numérique. Tous les autres lui échappent ou, tout au moins, ils ne peuvent prendre place dans les colonnes de nombres dont sont formées les publications statistiques. Aussi, ne peut-on faire la statistique des actions immorales, mais seulement celle des actions déclarées immorales par la loi, les infractions qui, parvenues à la connaissance de la justice, ont été l'objet d'une décision coulée en force de chose jugée (2).

---

(1) CLAUDE BERNARD. Introduction à l'étude de la médecine expérimentale (première partie, chap. I).

(2) Non seulement les actes immoraux, mais même les délits dont les auteurs sont restés inconnus, ne sont pas compris dans la statistique.

Le moraliste qui se livrerait à l'étude de la criminalité voudrait connaître toute l'étendue de ce phénomène ; non seulement la criminalité réprimée par les tribunaux, mais encore celle qui a échappé à la justice ; non seulement la criminalité légale, mais aussi celle qui se cache sous les dehors de la correction de la vie sociale. La statistique doit renoncer à cette ambition : elle ne connaît que les nombres, elle ne fait que compter.

L'investigation faite en vue de la statistique n'est donc pas une application complète de l'observation scientifique ; elle est limitée dans son objet et, dès lors, elle ne peut se confondre avec un procédé logique aussi général que l'observation proprement dite. C'est la raison pour laquelle un terme spécial, celui de *relevé,* convient mieux pour désigner le procédé statistique dans la phase où il s'applique à déterminer les caractères des phénomènes.

Nous définissons le relevé : « l'investigation, unité par unité, portant sur les manifestations individuelles d'un phénomène et les conditions de toutes ses manifestations ».

II. — Limites de l'application du procédé.

63. L'application de la méthode n'étant possible qu'à la condition que se trouvent réunies certaines conditions générales, il semble rationnel de consacrer d'abord quelques lignes à l'exposé de ces conditions.

En premier lieu, la recherche à faire ne peut heurter les sentiments intimes de ceux qu'elle concerne, elle ne doit pas faire naître parmi eux des appréhensions de nature à altérer la sincérité des réponses. L'oubli de ces mobiles d'ordre psychologique est de nature à altérer gravement la valeur du relevé statistique. Dans les recherches démographiques, les chances de heurter les sentiments intimes sont relativement peu nombreuses, les questions posées aux citoyens ne pouvant guère provoquer de susceptibilités : le sexe, l'état civil, l'âge, la profession... Certaines de-

mandes des recensements démographiques n'échappent cependant pas entièrement à cette cause d'erreur. Le nombre de personnes mariées ne serait-il pas plus grand d'après le recensement que d'après les relevés de l'état civil? Certaines personnes sans profession avouable n'en indiquent-elles pas une autre pour les besoins de la cause?

Lorsqu'on introduit dans un recensement des questions touchant au domaine de la conscience, on risque de faire fausse route; telle est la demande qui fut formulée dans l'un des derniers recensements italiens, à peu près en ces termes : « A quel culte appartenez-vous? » La question, a fait remarquer un auteur, est à double entente. On peut appartenir à un culte en ce sens qu'on est né dans ce culte, qu'on y a été élevé, qu'on en a même rempli des obligations essentielles, sans toutefois le pratiquer encore. S'il s'agissait de s'informer des pratiques religieuses *actuelles* des recensés, l'inconvénient était encore plus grave, car la question, en ce cas, aurait été directement à l'encontre de la règle qui place en dehors de la statistique le domaine de la conscience. Dans les recherches relatives aux choses de la vie économique, apparaît souvent le danger de heurter trop directement, par des questions indiscrètes, le sentiment de l'intérêt personnel. Déjà en 1846, Quetelet écrivait ces lignes, encore vraies aujourd'hui : « Si l'on croit reconnaître dans des questions un but fiscal ou une curiosité inquisitoriale, l'on ne se fera pas de scrupule de donner des réponses inexactes (1). »

Il n'est pas toujours possible d'éviter complètement que certaines investigations ne paraissent aller à l'encontre de l'intérêt personnel ou collectif des recensés. Dans ce cas, on s'efforce, au moyen d'une revision attentive du matériel statistique, de faire disparaître, ou d'atténuer dans une

(1) QUETELET. *Lettres sur la théorie des probabilités*, p. 290.

mesure notable, les sources d'erreurs possibles : nous parlerons de ces procédés lorsque nous examinerons les moyens employés par la critique statistique. Bornons-nous à retenir ici cette règle essentielle : dans une enquête statistique, il y a un danger réel à heurter de front des défiances profondément ancrées dans l'esprit du public, ou à employer des moyens d'investigation, dont la nature même ferait naître de telles défiances. Organiser auprès des chefs d'entreprise une enquête statistique sur les heures de travail des ouvriers adultes, en faisant recueillir les données par les inspecteurs du travail, revient à peu près à affirmer les sympathies du gouvernement à l'égard d'une extension de la réglementation de la durée du travail à une catégorie de personnes non assujetties encore à de pareilles restrictions. La sincérité des réponses données par les patrons pourrait, à bon droit, paraître douteuse.

64. A ces causes, d'ordre psychologique, limitant la portée du relevé, s'ajoutent des raisons d'ordre administratif et pratique.

Le temps à consacrer à une recherche déterminée peut paraître trop long par rapport à l'intérêt qu'elle présente. Dans le domaine scientifique pur, où les recherches sont conduites par des savants, sous leur responsabilité personnelle, il n'y a guère de place pour des scrupules de ce genre ; en cette matière, rien de ce qui ajoute une parcelle de vérité aux connaissances déjà acquises ne peut sembler inutile ; des savants consacrent parfois toute une vie à élucider l'un ou l'autre point obscur, sans se préoccuper de l'importance relative qu'il peut présenter. C'est au point de vue de l'absolu, celui de la science, qu'ils se placent. Si l'on avait dû aborder les problèmes scientifiques avec un esprit « pratique », nos connaissances générales — cela est certain même du point de vue des applications — seraient autrement limitées qu'elles ne le sont aujourd'hui. Mais la statistique est surtout utilisée pour

les fins de l'administration ; il est rationnel d'envisager ses recherches sous l'aspect pratique : la relation entre la peine et le résultat attendu.

L'Etat, principal organisateur des grandes enquêtes statistiques, peut reculer devant la dépense importante qu'exige une recherche étendue. A plus forte raison, les sociétés savantes, les institutions privées, les particuliers, sont-ils en droit d'hésiter à entreprendre des investigations longues et coûteuses. Le prix de revient des travaux statistiques a contribué à créer en faveur de l'Etat une sorte de monopole. Cela est fâcheux, parce que les gouvernements sont rarement bien disposés en faveur des recherches scientifiques et qu'ils ont grand besoin d'être aiguillonnés par la concurrence. Pour se faire une idée du coût d'un travail statistique important, il convient de prendre pour exemple un grand pays, comme les Etats-Unis, et d'autre part une vaste opération, un recensement général. Le tableau ci-après servira d'illustration à cet égard.

Personnel employé au recensement et coût du travail aux États-Unis, de 1830 à 1900 (1)

Recensements.	Nombre d'agents du relevé.		Nombre maximum d'agents employés à la tabulation.	Population totale.	Coût total.	Coût par tête.
	Chefs	Recenseurs			$	$
1830	36	1,519	43	11,866,020	378,545	0 0294
1840	41	2,167	28	17,069,453	833,370	0.0488
1850	45	3,231	160	23,191.876	1,423,350	0 0613
1860	64	4,417	184	31,443,321	1,969,376	0.0626
1870	75	68 (2) 6,330	438	38,558,371	3,421,198	0.0877
1880	150	31,382	1495	50,429,345	5,790,678	0 1148
1890	175	46,804	3143	62,979,766	11,547,127	0 1833
1900	300	2,648 (2) 52,871	3554	76,149,386	11,854,817	0 1550

(1) Extrait de *American Census taking*, publication du Département du Commerce et du Travail, 1903 (a paru originairement dans *The century magazine*, avril 1903).

(2) Agents employés à la campagne, distincts des agents recenseurs.

Malgré la dépense énorme que représente la somme de 11,854,817 dollars qui fut consacrée au recensement américain de 1900, aucun employé ne fut payé pour son travail au delà de ce que gagne un ouvrier intelligent. Le fait peut surprendre, mais quand on sait la somme incalculable de travail que représente un vaste recensement, on comprend que, même avec un barême de salaires modestes, la dépense soit aussi grande.

Il ne faut pas s'attendre à voir diminuer le prix de revient des opérations statistiques : leur programme, de plus en plus compliqué, rendra nécessaire pour son exécution le concours de compétences mieux établies encore que par le passé, et les exigences en matière de rémunération et de salaires ne feront sans doute qu'aller en augmentant.

Les limites du relevé n'ont rien de fixe ou d'immuable. Plus rigides dans les pays arriérés, elles seront plus souples parmi ceux avancés en civilisation. Dans un même pays, elles sont susceptibles de reculer progressivement à mesure que les particuliers verront s'évanouir leurs défiances, que l'organisation statistique sera mieux comprise, que l'Etat attachera plus d'importance aux recherches statistiques.

III. — Divisions du relevé direct.

65. Le relevé est de deux espèces : le relevé statistique ou *direct* et le relevé *indirect*.

Le relevé direct est celui qui consiste dans l'énumération de toutes les manifestations individuelles du phénomène que l'on étudie. Il a pour type le recensement.

L'étendue du relevé direct n'est pas toujours aussi générale que la définition donnée plus haut porterait à le croire. Des nécessités pratiques conduisent parfois à restreindre la définition des unités que l'on compte. Ainsi la statistique criminelle ne s'étend qu'aux crimes et délits dont les

auteurs ont été découverts; la statistique des accidents du travail (en Belgique) ne comprend que les accidents survenus dans les entreprises assurées par les institutions agréées. En Angleterre, la statistique des grèves ne relève pas les conflits qui se sont produits dans des entreprises employant moins de cinq ouvriers, etc.

Certaines unités peuvent donc être omises quand leur importance est minime.

On se trouve encore dans le domaine du relevé direct quand toutes les unités répondant à une définition générale sont dénombrées. Bien qu'un recensement industriel ne vise le plus souvent que les entreprises industrielles privées, à l'exclusion de celles de l'Etat, des provinces et des communes, malgré que certaines industries se trouvant aux confins de l'agriculture ou du commerce puissent être négligées, un tel recensement, même limité, appartient au relevé direct parce qu'il s'étend à toutes les entreprises industrielles visées par une définition générale. Il n'en serait pas de la sorte si le nombre des industries à recenser se trouvait limité à certaines catégories ou groupes; il s'agirait alors d'un relevé partiel.

66. Le relevé direct se divise en :

a) *Relevé continu* relatif à des phénomènes se présentant sans interruption, observés selon certaines divisions du temps. Se trouvent dans ce cas les statistiques se rapportant au mouvement de la population (naissances, décès, mariages, divorces), aux importations et exportations de marchandises, à la criminalité, au chômage, aux grèves, aux accidents du travail, etc. Ces statistiques supposent que le relevé s'exécute d'une manière ininterrompue, d'après des règles fixées par une organisation générale;

b) *Relevé périodique,* qui vise certains phénomènes, ne se modifiant que lentement et pouvant être observés de temps

à autre. Les recensements sont le type de ce genre de recherche. La population d'un pays ne se modifie que lentement; on se contente donc de la compter tous les cinq ou tous les dix ans. Le nombre d'entreprises industrielles, leur forme d'exploitation, la répartition des entreprises d'après le nombre d'ouvriers qu'elles occupent, les salaires payés aux ouvriers et le nombre d'heures de travail exigées en échange sont sujets à des modifications incessantes, mais il suffit d'en observer les résultats à des moments plus ou moins espacés; aussi les recensements industriels ne se font-ils que tous les dix ans, parfois même la périodicité du relevé n'est pas fixe;

c) *Relevé occasionnel,* c'est-à-dire effectué seulement lors de certaines circonstances exceptionnelles (maladies contagieuses, établissement des bases d'un impôt, etc.).

Il est à remarquer que ces divisions du relevé ne précisent sa nature que pour autant qu'elle est caractérisée par la continuité ou le renouvellement de l'investigation. La manière dont le relevé est accompli donne lieu aux deux divisions ci-après :

67. Le relevé direct peut s'exécuter de deux façons :

a) *Relevé automatique :* c'est celui qui s'effectue lui-même par le libre jeu des institutions ou des coutumes.

Dans la plupart des pays, le mariage n'est reconnu par la loi que s'il a été contracté devant un officier de l'état civil chargé d'en dresser acte sur le champ. D'après le Code civil, les actes de naissance doivent être rédigés de suite après la déclaration de naissance qui doit être faite par le père dans les trois jours de l'accouchement. Ces prescriptions, de même que celles relatives aux décès, organisent un véritable relevé automatique des actes les plus importants de la vie civile. Le relevé automatique présente le plus de sécurité et, chaque fois que les statisticiens peuvent y recourir, ils ne manquent pas de le faire. Toutefois, il faut s'en

tendre sur le caractère spécial de ce relevé. Comme il est organisé par l'autorité administrative, ou qu'il est la suite d'une prescription législative quelconque, il n'embrasse pas toujours la totalité des manifestations individuelles du phénomène visé, mais seulement celles qui répondent à une définition spéciale. Les définitions juridiques, notamment, ne correspondent pas toujours à toute la réalité. Si l'on veut substituer un point de vue plus général à celui-là, on se trouvera devant une difficulté : celle de compléter les données déjà recueillies. On recule le plus souvent devant cette tâche. Une précaution élémentaire s'impose alors : il faut délimiter avec le plus grand soin le champ couvert par la statistique.

b) Le relevé est dit *réfléchi* ou *voulu* quand il faut rechercher spécialement les unités en vue de les dénombrer. Ce relevé exige beaucoup de soin, tant pour éviter les répétitions ou doubles emplois, que pour prévenir les omissions.

CHAPITRE II

Organisation du relevé statistique.

I. — Préparation du relevé.

68. La préparation du relevé est chose d'importance. Il est impossible d'y apporter trop de soin et de réflexion, la technique, en cette matière, étant une affaire capitale. Trop de théoriciens sont enclins à considérer ce problème comme d'une importance secondaire, ou sont portés à croire que toutes les difficultés se trouvent aujourd'hui résolues. Il semble que l'attention des auteurs les plus récents se porte du côté de la mise en œuvre du matériel statistique de préférence à ce qui regarde le relevé. Nous serait-il per-

mis de trouver qu'il y a là une erreur? La technique du relevé n'est pas parfaite, même après tant d'expériences et un certain nombre de désillusions. Il n'est pas exact qu'on puisse se désintéresser de tout ce qui la concerne comme si l'on se trouvait en face d'un problème depuis longtemps résolu. De nombreux progrès sont à réaliser avant d'arriver à une solution satisfaisante. Le soin qu'apportent les biologistes à fixer la technique de leurs recherches contient un enseignement dont les statisticiens doivent faire leur profit. Non seulement les principes généraux sont déterminés avec un soin extrême — un simple « manuel » de technique microscopique est un gros volume (1), — mais, de plus, il existe dans tous les pays des périodiques où ne paraissent que des articles relatifs aux perfectionnements qui se trouvent, de jour en jour, apportés aux méthodes de recherche. Il n'y a presque pas de mémoire qui se publie sans que l'auteur, non seulement expose et discute sa technique, mais encore indique l'une ou l'autre amélioration qu'il a apportée aux procédés usités avant lui. Dans les laboratoires, la formation scientifique est tout d'abord orientée vers la technique, les précautions les plus minutieuses sont réunies pour que l'observation se fasse dans des conditions irréprochables ; dans les grandes institutions scientifiques, tout, jusqu'à l'atmosphère de recueillement qui les enveloppe, jusqu'à leurs longs corridors ouatés de silence, tout parle de recherches méticuleuses, patientes, orientées avec ferveur vers la découverte du vrai.

Nous n'en sommes pas là en statistique.

Programmes mal établis, tantôt trop étendus, tantôt trop étroitement circonscrits — questions indiscrètes ou à double entente — disposition typographique défectueuse des questionnaires utilisés, agents recenseurs mal choisis, peu prépa-

---

(1) *Le Manuel de technique microscopique*, de M. E. de Rouville (1913), ne comporte pas moins de 720 pages in-8°, et ce n'est qu'un « court résumé » des méthodes les plus usuelles, avec renvoi aux mémoires originaux.

rés à leur mission, mal payés, — travail hâtif de mise en œuvre — tableaux d'où ne se dégage pas un rayon lumineux qui mette sur la piste d'un fait important le chercheur découragé: ce sont là défauts trop fréquents et sur lesquels, maintes fois, au cours de cet exposé, nous aurons à revenir. A côté d'œuvres habilement agencées, il n'en est que trop laissant à désirer. Depuis quelques années, les aperçus techniques se font plus fréquents et les grandes administrations montrent moins de discrétion dans l'exposé des méthodes à l'aide desquelles les données statistiques ont été recueillies et mises en œuvre. Dans un travail scientifique, rien ne vaut la sincérité; une erreur, franchement reconnue, est presque aussi profitable à la science qu'une vérité démontrée. Les améliorations constatées sont dues, en partie, à l'action de l'Institut international de statistique, à qui on ne saurait en être trop reconnaissant. C'est par la critique comparée des méthodes que la statistique se perfectionnera, comme se sont perfectionnées les sciences naturelles par l'usage d'une technique de plus en plus sûre, mise au service de l'hypothèse scientifique.

69. Comme dans tout travail, il faut, avant d'entamer une recherche statistique, avoir un plan, se tracer un programme logiquement ordonné. Les qualités requises pour ce programme ne sont pas différentes de celles recommandées dans tout plan quelconque, mais elles s'étendent à quelques matières qu'on n'est pas dans l'habitude de considérer dans les études économiques ou sociales ordinaires; c'est la raison pour laquelle cette question si simple de l'élaboration du plan ou programme demande à être examinée avec quelque détail.

Nous remarquons, en premier lieu, que l'élaboration du programme se fera dans des conditions très différentes, selon que l'on aura à organiser un relevé fait d'après des documents recueillis en vertu de prescriptions spéciales

(relevé automatique) ou un relevé de personnes ou de faits n'ayant encore été l'objet d'aucune observation (relevé réfléchi).

Dans le premier cas, l'auteur du programme se trouve nécessairement devant un champ de recherches limité ; il ne peut songer à faire porter des investigations sur un nombre de faits supérieur à ceux consignés dans les documents à utiliser ; il lui est impossible aussi d'en modifier la nature, et si ces faits sont amassés d'après un critère juridique, la recherche statistique se trouve limitée aux faits répondant à la définition juridique, sans plus.

La statistique des accidents du travail, par exemple, n'est pas le relevé de tous les accidents survenus au cours du travail, mais seulement de ceux qui répondent à la notion juridique d'accident de travail et se sont produits dans des entreprises soumises à la loi.

Si l'auteur du programme est chargé d'organiser lui-même la recherche, il jouit évidemment d'une plus grande liberté, sauf ce qui est dit plus haut relativement aux limites de la statistique.

Un plan ou programme bien ordonné comprendra la détermination précise de l'objet du relevé, l'étude du coût de l'opération, un essai de présentation des données et, dans les cas compliqués et nouveaux, des dispositions pour procéder à un essai de dénombrement. De plus, il fixera l'époque du relevé et les délais prévus pour son achèvement, l'unité géographique à laquelle les faits sont rapportés, les voies et moyens d'exécution, les causes d'erreur et les moyens de les faire disparaître.

Cette suite complexe de recherches comporte une prévision à longue portée. Celui qui organise un travail statistique ne peut se contenter de vivre au jour le jour et de parer aux difficultés au fur et à mesure qu'elles se présentent. Il lui faut des vues larges et précises sur toutes les opérations qui se dérouleront successivement, sinon à tout

instant il se trouvera devant le fait accompli sans possibilité de revenir sur ses pas lorsqu'une erreur lui apparaîtra. On ne recommence pas un relevé; on ne refait pas même un dépouillement lorsqu'il s'agit d'un travail portant sur quelques centaines de milliers ou quelques millions d'unités. La sûreté de vues ici n'est pas seulement une qualité désirable, c'est une exigence à laquelle nul ne peut se soustraire.

70. Ajoutons encore, ce qui est essentiel, que la rédaction du programme exige souvent chez l'auteur du document l'existence d'une hypothèse scientifique à vérifier. Sans doute, beaucoup de recherches statistiques se font aujourd'hui sur des bases consacrées par l'usage et par la science. Il est difficile d'innover en matière de statistique de la population, bien que, même dans ce domaine depuis si longtemps exploré, des points de vue nouveaux surgissent encore de temps à autre. Mais qui ne voit le rôle essentiel de l'hypothèse scientifique dans les questions où interviennent les facteurs psychologiques, moraux, économiques? Que sera la statistique criminelle si elle n'est guidée par l'hypothèse scientifique? Un amas confus de faits contradictoires sur lesquels le chercheur sera impuissant à projeter la moindre clarté. La statistique industrielle, celle des salaires et des heures de travail, celle des grèves, pour citer quelques exemples, ne peuvent donner de résultats vraiment intéressants que si les recherches, au moment de l'élaboration du programme, sont orientées en tenant compte des théories économiques.

Qu'on ne confonde pas l'usage de l'hypothèse scientifique avec le parti pris ou la partialité : en partant d'une théorie, le savant ne la considère que comme une assertion soumise à la vérification; c'est une orientation donnée aux recherches, rien de plus, et les faits seront par lui recueillis avec autant de soin, acceptés avec autant de respect s'ils détruisent la thèse que s'ils la confirment.

L'usage de l'hypothèse scientifique imprime au programme et à ses divisions une qualité maîtresse : la clarté. Tout se trouvant ordonné selon une idée directrice, les différentes parties se trouvent facilement disposées dans leur ordre logique, les questions deviennent précises, elles se complètent l'une l'autre, les détails dans lesquels on entre ne sont pas oiseux, il n'y a pas de demandes inutiles.

71. Avant tout, il importe de préciser les caractères de la chose que l'on compte : l'*unité*. Celle-ci forme la base de la recherche statistique; on en compte le nombre d'abord; on en relève ensuite les qualités. Dans un recensement de la population, l'unité est l'individu; celui-ci peut être considéré à deux points de vue : l'individu présent, au moment du recensement, dans un lieu donné; l'individu qui a sa résidence habituelle dans un lieu donné, mais qui peut en être momentanément absent. Le premier point de vue donne la population de fait; le second permet de déterminer la population de droit. On peut adopter l'un ou l'autre, ou, ce qui est plus fréquent, tous les deux à la fois. Lorsque cette question est décidée vient la définition des qualités de l'individu : il peut être un homme ou une femme (répartition par sexe), jeune ou âgé (répartition par âge), marié ou non (répartition par état civil), exerçant une profession ou n'en exerçant aucune (répartition par profession ou condition), habiter la ville ou la campagne (répartition géographique).

Dans toutes ces hypothèses, l'unité reste la même, c'est elle qui donne à la statistique sa physionomie propre.

La statistique criminelle emploie une double unité : l'infraction ou le délit et le délinquant. La statistique de l'industrie et du commerce d'abord a pour unité l'entreprise industrielle ou commerciale; elle compte les entreprises, puis elle analyse leur caractère : l'industrie à laquelle elles appartiennent, leur état d'activité ou de chômage, leur

situation géographique, leur importance, le mode d'exploitation duquel elles relèvent, leur grandeur, etc. On peut dire aussi que la statistique industrielle a une seconde unité : l'individu occupé dans l'entreprise industrielle. Après le comptage de ces unités, vient l'étude des particularités qu'elles présentent : le sexe, l'âge, la situation sociale, le salaire, les heures de travail, la résidence dans une localité autre que celle où l'individu trouve son occupation.

72. Pour pouvoir former la base de la statistique, l'unité doit réunir certaines qualités; on peut les résumer de la sorte :

1° Elle doit appartenir à l'ordre des phénomènes collectifs, sinon il existe des procédés plus rapides et moins coûteux que la méthode statistique pour en déterminer la nature et les caractères;

2° Elle doit être susceptible de se réduire en une expression numérique. Il faut donc que les objets qui constituent les unités se présentent par unité individuelle : une maison, un ménage, une famille. C'est pour cela qu'on dit que le relevé est l'observation des manifestations individuelles d'un phénomène. Si les faits se présentaient en masse confuse, on ne saurait compter les unités dont ils se composent, donc, on ne pourrait en faire la statistique.

C'est une erreur de vouloir comprendre dans des statistiques des choses qui ne se différencient que par leurs qualités (bon, mauvais, grand, petit, etc.), car ceci est affaire d'appréciation subjective. Dans certaines recherches, on substitue des chiffres aux mots qui expriment des qualités : ainsi, à une récolte qu'on espère devoir être très bonne, on attribuera la cote 75, à une bonne la cote 50, à une mauvaise la cote 25. Les chiffres ne changent rien à l'affaire, il s'agit toujours d'une appréciation. Le procédé dont il s'agit n'est d'ailleurs employé que dans l'estimation

de la récolte espérée et nous verrons plus loin que les estimations dépendent du relevé indirect.

3° Elle sera de nature à pouvoir être reconnue d'une façon qui exclut tout doute. Nous reviendrons sur ce point très important à l'endroit où nous parlerons de la rédaction du questionnaire. Pour que l'unité soit telle qu'elle se distingue facilement de toutes les autres, il n'est pas nécessaire que sa nature soit différente des autres choses ; ce serait limiter, d'une façon arbitraire, le domaine de la statistique ; il suffit qu'une bonne définition en soit donnée. La définition peut être vulgaire — comme celle donnée par le dictionnaire — ou juridique, ou économique ; parfois, les notions fournies par ces définitions sont encore trop peu précises par rapport au but visé et l'on doit s'arrêter à une définition statistique spéciale.

73. La définition vulgaire prête à de nombreux malentendus ; souvent, elle définit les mots par les mots et n'apprend rien de précis. M. Bertillon a donné un exemple typique des divergences auxquelles on arrive en empruntant, sans modification, des définitions à la langue vulgaire : ayant à dresser simultanément un relevé des maisons de Paris, l'administration municipale et le ministère des finances arrivèrent à des résultats fortement discordants. Comment est-il possible de se tromper aussi grossièrement dans le comptage des maisons ? se demandèrent les sceptiques. La cause de la divergence était simple : l'administration municipale avait compté les habitations, le fisc les locaux imposables. Les deux statistiques poursuivant des buts distincts, il n'y avait pas, dans ce cas, d'erreur à proprement parler, mais l'emploi du terme « maison » prête à confusion et une erreur de ce genre aurait pu être commise dans toute statistique qui se serait bornée à employer ce mot sans l'accompagner de commentaires.

On a confondu maintes fois, sous l'expression de mort-

nés, les enfants sans vie et les enfants non viables décédés avant la fin du troisième jour qui suit leur naissance. Dans quelques pays, la distinction était établie, ailleurs elle ne l'était pas et ce défaut de concordance fut longtemps une cause de trouble dans la comparaison des statistiques.

Les définitions juridiques sont, en général, plus précises que la définition vulgaire. Quelquefois, cette précision est un défaut, en ce sens que la définition légale vise un cas particulier, est rédigée dans un but spécial, ce qui en exclut des cas tout aussi intéressants que la statistique a intérêt à connaître. Aux termes de la loi italienne, du 31 janvier 1901 sur l'émigration, loi qui a pour but de prendre des mesures de protection en faveur des émigrants, « doivent être considérés comme émigrants tous ceux qui, voyageant par mer en troisième classe, se dirigent vers des pays situés au delà du détroit de Gibraltar ou, s'ils sont au nombre de plus de cinquante, vers des pays au delà du canal de Suez ». Cette définition, très spéciale, ne peut aucunement servir à la statistique de l'émigration : celle-ci n'a pas à se baser sur le mode de transport, ni sur la classe du billet du voyageur.

La statistique du chômage fournit un exemple d'une définition économique. Chômage s'oppose à activité; dans un sens large, un chômeur est quelqu'un qui ne travaille pas. Mais faudrait-il compter parmi les chômeurs les personnes sans profession? Ce serait absurde. Faut-il considérer comme chômeurs les malades, les invalides, les infirmes, les vieillards? La chose est très contestable. Si l'on met en rapport le nombre de chômeurs et l'activité industrielle, il vaut mieux, semble-t-il, définir comme chômeur l'ouvrier qui, en état de travailler et désirant travailler, est sans ouvrage. Une définition de ce genre est essentiellement propre à la science économique.

Les définitions purement statistiques sont fort nombreuses. En voici une qui est adoptée par la statistique

internationale : c'est la définition de l'unité ouvrière en matière de statistique des accidents du travail. Pour déterminer la fréquence des accidents, il faut pouvoir apprécier d'une manière aussi exacte que possible le nombre des ouvriers soumis au risque. Or, le nombre des ouvriers occupés dans un établissement industriel varie souvent au cours de l'année. Devant la difficulté, voire l'impossibilité d'obtenir à cet égard des données exactes, les Congrès internationaux ont préconisé d'adopter une mesure commune. Le nombre de journées de travail effectuées par l'ensemble du personnel ouvrier peut toujours être connu avec exactitude et ce nombre de journées représente les journées effectives de travail, y compris les fractions de « journées » prestées par le personnel au delà de la durée normale du travail. Dès lors, si l'on adopte comme diviseur de ce nombre un chiffre représentant la quantité de jours d'exploitation, on peut arriver à une certaine grandeur idéale qui exprime le nombre de travailleurs complets, c'est-à-dire de travailleurs ayant tous été occupés pendant le même temps. L'Autriche et l'Allemagne ont les premières adopté le nombre de 300 jours d'exploitation et ont donné le nom de *Vollarbeiter* au quotient de la division des journées effectives par le chiffre 300. La même unité a été adoptée par la statistique belge des accidents du travail (1).

II. — Le relevé considéré sous le point de vue du temps.

74. Tous les phénomènes quelconques peuvent être considérés sous l'aspect du temps et de l'espace. Essayons d'indiquer de quelle manière ces notions générales conditionnent et limitent le relevé et les phénomènes auxquels il s'applique.

---

(1) Royaume de Belgique. Office du travail. *Statistique des accidents du travail*, t. II, p. 576. Bruxelles, 1912.

Sous le rapport du temps, le relevé peut être considéré sous trois aspects :

1° quant à la durée de l'observation;

2° quant à l'époque de l'observation;

3° quant à la durée du phénomène.

En ce qui regarde la durée de l'observation, il faut distinguer entre le relevé qui vise à reproduire l'état d'un phénomène, de celui qui cherche à en faire connaître le mouvement. Dans le premier cas, on essaye de reproduire les faits dans des conditions telles qu'ils échappent aux accidents de variabilité. Dans le second, on s'efforce au contraire de faire apparaître les conditions diverses influençant le phénomène, de façon à connaître les causes générales et les causes accidentelles. En démographie, un recensement de la population forme un type de statistique du premier genre; le relevé des naissances, des décès, des mariages, appartient à la seconde catégorie, dont l'objet est le mouvement de la population.

Examinons d'abord les faits dans leurs rapports avec la durée du relevé.

Le synchronisme des opérations du relevé est, dans ce cas, une condition essentielle. Le fait à enregistrer est-il simple, n'y a-t-il que peu d'unités à relever, les difficultés à vaincre sont de minime importance et il n'y a pas lieu de s'y arrêter ici. Au contraire, lorsque le fait est complexe, lorsqu'il se compose d'une multitude de manifestations individuelles, le statisticien doit prendre des précautions minutieuses afin d'éviter de mêler les faits d'hier aux faits d'aujourd'hui. Il n'y aurait aucune précision dans un relevé confondant dans un ensemble hétérogène, les faits passés avec les faits actuels. Pendant l'enregistrement de certains faits, ceux-là déjà compris dans le relevé auraient changé d'aspect. On sent trop bien que le relevé de l'*état* d'un phénomène doit être synchronique pour devoir s'at-

tarder à démontrer la nécessité de la simultanéité de toutes ses parties : ce serait faire la démonstration de l'évidence même. A un point de vue idéal, une statistique de « situation » devrait être une photographie instantanée. En pratique, on se trouve encore fort éloigné de la perfection. Celle-ci est d'ailleurs hors de notre portée : admettons la possibilité de recenser la population en quelques heures; pendant ce temps peuvent survenir des décès et des naissances non consignés sur les bulletins.

Notre ambition se limite donc à raccourcir, par une bonne organisation administrative, le délai exigé pour les réponses à donner, et à veiller avec une attention scrupuleuse au synchronisme de toutes les parties du relevé (1).

75. Alors que dans les observations relatives à l' « état » d'un phénomène on s'efforce de rendre aussi brève que possible la durée du relevé, on tend, au contraire, à la prolonger autant qu'on le peut quand il s'agit de faire appa-

(1) Les mesures prises pour atteindre ce but sont parfois d'une grande ingéniosité. Celles qui ont été mises en vigueur par le gouvernement britannique aux Indes pour le recensement de la péninsule méritent qu'on s'y arrête, car elles sont réellement instructives. Le dernier recensement de l'Inde, qui date du 30 mars 1911, a été, comme les dénombrements précédents, accompli en une seule nuit, avec le concours de plus de deux millions de recenseurs. Tout le territoire de l'Inde se trouve divisé en « blocs », comprenant chacun de trente à cinquante habitations; au-dessus vient le « cercle », comprenant de dix à cinquante « blocs ». Les « cercles » sont réunis d'après les divisions administratives (tahsils, taluks, etc.). Enfin, la centralisation suprême est confiée à un commissaire spécial pour le recensement de l'Inde. Quelques semaines avant la date fixée pour le recensement, les agents recenseurs font le tour des habitations qui leur sont désignées et consignent sur le bulletin les réponses concernant toutes les personnes qui ont à cet endroit leur résidence habituelle. Lorsque la nuit du recensement est arrivée, les agents se présentent à nouveau, biffent les indications du bulletin relatives aux personnes non présentes et portent sur ce document les réponses concernant les nouveaux venus. Cette mission, confiée à des agents indigènes, est, à ce qu'on assure, remplie avec un soin extrême, au point que les réponses libellées par les Européens sont inférieures, comme précision, à celles formulées par les agents hindous. Le synchronisme est donc réalisé d'une façon très satisfaisante dans l'ensemble; il y a, naturellement, des exceptions motivées par des maladies

raître le « mouvement » des faits étudiés. La théorie pure élèverait ici une prétention aussi irréalisable que celle qui consiste à faire du relevé une photographie instantanée: ce serait de réunir des observations portant sur le temps entier pendant lequel le phénomène a continué à se manifester. Autant assigner à la statistique une durée qui se confondrait avec celle de la vie de l'humanité. Sans nous arrêter à des rêves de cette nature, reconnaissons l'évidence; l'aspect toujours changeant des faits observés par la statistique exige une longue période de temps couverte par l'observation, pour que l'on ait chance de découvrir les lois permanentes qui gouvernent les faits de la nature. C'est seulement lorsque les faits amassés sont assez nombreux, que nous nous trouvons dans les conditions voulues pour observer leur régularité; or, leur répétition étant limitée dans le cours d'une année, par exemple, il faut, pour réunir des cas assez nombreux, que le relevé couvre une série d'années. Les changements d'une année à l'autre peuvent être dus à des causes accidentelles; c'est seulement à condition d'étendre la période d'observation qu'on verra se dégager les tendances plus profondes au-dessus des variations de courte durée. La statistique morale a surtout besoin de cette observation prolongée, qui, pour être réalisée dans des con-

---

contagieuses, des troubles, des événements imprévus, etc. Des difficultés, propres à la péninsule, sont celles résultant de la présence dans les trains ou dans les steamers de voyageurs pendant un parcours qui exige souvent plusieurs jours, du travail de bûcherons qui s'enfoncent dans les forêts et y séjournent plusieurs semaines, de l'existence de nombreux endroits consacrés où se rendent des milliers de pèlerins. Pour recenser les voyageurs, on a imaginé les dispositions suivantes : toute personne qui prend le train après 7 heures du soir la nuit du recensement, est recensée sur le quai d'embarquement, si on en a le temps, sinon dans le train même. Les voyageurs qui montent dans le train au cours de la nuit sont recensés de même, à moins qu'ils n'exhibent un certificat établissant qu'ils ont déjà été recensés. Vers 6 heures du matin, tous les trains s'arrêtent et on procède à la visite des wagons pour s'assurer si personne n'a été omis du recensement (*). La population de l'Inde était, le 30 mars 1911, de 315,156,396 habitants.

(*) *Census of India* 1911, vol. I. Report by E. A. Gait, Introduction. Calcutta, 1913.

ditions convenables, exige le relevé automatique ou continu.

Il ne faut pas confondre la durée des observations avec l'unité de temps utilisée pour la présentation. La statistique des décès est organisée de telle façon qu'on puisse la présenter d'une manière très détaillée quant au temps où les décès sont enregistrés ; parce que cette statistique demande à être suivie pendant longtemps pour pouvoir déterminer les coefficients de mortalité, suit-il de là qu'on puisse se borner à relever et à publier les décès à de longs intervalles, par exemple pour une année entière? Il n'y a aucun rapport entre les deux ordres d'idées. Les observations les plus longues peuvent gagner beaucoup à être présentées à intervalles fort rapprochés ; il en est ainsi des naissances et des décès pour lesquels on remarque de grandes variations par mois, et même par jour et par heure. Les statistiques météorologiques seraient dénuées d'intérêt si on se bornait à en publier les résultats par mois ; il faut descendre, au contraire, dans le détail pour en tirer les enseignements qu'elles comportent. A cet égard, les appareils enregistreurs rendent les plus grands services à l'aide des graphiques qu'ils tracent automatiquement.

76. Le relevé doit donc satisfaire à deux conditions : être le plus prolongé possible, être le plus particularisé qu'il se peut. Ces conditions ne sont pas contradictoires, mais complémentaires. Dans le cas du relevé continu comme du relevé périodique, l'observation isolée sera soustraite aux modifications successives des évènements ; elle sera rapide et, si c'est nécessaire, répétée. Mais dans les statistiques du « mouvement » social, le temps sur lequel portera l'observation sera aussi long qu'on le pourra. Au point de vue de la durée, on ne peut mieux comparer les statistiques d'état et de mouvement qu'à la photographie instantanée et à la cinématographie. Un film est une succession ininterrompue de photographies, prises très rapidement ; si l'on pouvait relever les phénomènes de la vie sociale avec une égale pré-

cision, on posséderait des statistiques parfaites. Le procédé, malheureusement, ne s'applique pas à tous les faits qu'il serait intéressant de connaître. Il faut alors se borner à prendre, à un moment opportun, une photographie instantanée des phénomènes, avec les précautions voulues pour assurer un synchronisme parfait.

Quels sont les phénomènes à observer par le moyen du relevé continu? Quels sont ceux auxquels convient le relevé périodique ou occasionnel? La réponse se trouve déjà donnée dans la définition des modalités du relevé, nous y revenons sommairement en conclusion de ce qui précède. Parmi les faits soumis au relevé statistique, il en est dont l'évolution relativement lente exige seulement un certain nombre de points de repère suffisamment espacés; entre ces points, on pourrait noter des changements, mais ils sont peu importants dans l'ensemble. Un voyageur qui circule sur une route compte les bornes kilométriques, placées bien en évidence le long du chemin; il ne cherche pas du regard les petites bornes hectométriques à moitié dissimulées sous l'herbe des talus et des fossés.

La population d'un pays se modifie constamment; le changement, par rapport à la masse, en une semaine, un mois, une année a-t-il une réelle importance? Entrées et sorties se compensent, ce n'est qu'après un intervalle assez long que les diminutions ou les augmentations successives parviennent à apporter au total un changement assez sensible pour être noté. Ce qui est vrai du nombre des habitants, l'est encore bien davantage du rapport des sexes, de la répartition des âges, de la population des villes et des campagnes, de l'importance relative des professions industrielles et agricoles, de la proportion des personnes actives et des personnes sans profession. Il est bien suffisant de vérifier tous les dix ans si les faits relevés lors du précédent recensement sont restés les mêmes ou non.

Même lorsqu'une période est caractérisée par un mouvement économique intense, les salaires, dans l'ensemble, ne haussent pas assez vite pour qu'il soit nécessaire d'opérer des relevés périodiques à court intervalle; on peut en dire autant des heures de travail. En Belgique, bien que les années 1896 à 1901 aient été marquées d'une grande activité, la statistique des salaires dans les industries textiles, dressée au mois d'octobre 1901, n'a montré que peu de changements avec celle faite le même mois en 1896 : dans l'ensemble des industries textiles, les salaires n'ont guère varié ni pour les hommes, ni pour les femmes, ni pour les enfants ; pour les salaires des hommes et ceux des femmes, il y a lieu de noter un léger mouvement d'ascension des taux les plus bas vers les taux plus élevés, toutefois ce mouvement n'atteint pas, pour les hommes, un pour cent du nombre des ouvriers gagnant le même taux de salaire en 1896 (1).

Il nous suffit de connaître la distribution des salaires ou le chiffre de la population un jour donné, étant donnée la fixité relative de ces données pendant une période assez longue. Mais pourrions-nous nous déclarer satisfaits de savoir le nombre des décès et des naissances, les causes de décès un seul jour de l'année? Il est évident que non. Nous sommes là en présence de faits dont l'équilibre ne s'établit que lentement, le relevé doit être continu pour pouvoir grouper des données utilisables.

La base de la distinction à établir doit donc être cherchée dans les caractères de variabilité plus ou moins accentuée que présentent les phénomènes.

77. Pour les statistiques qui se font à un moment donné, il y a lieu de se demander quel est le moment le plus favorable. Affaire de bon sens et d'expérience; le choix de

---

(1) Royaume de Belgique. Office du travail. *Salaires et durée du travail dans les industries textiles au mois d'octobre 1901*, p. 301. Bruxelles, 1905.

l'époque ne se fixe pas d'après des règles théoriques. Pour certains relevés statistiques, telle époque convient; pour d'autres la date favorable à une exécution convenable des opérations sera différente. Si l'on veut recenser la population industrielle d'un pays, on ne choisira pas une date à laquelle de nombreux émigrants saisonniers se trouvent absents, ni une période durant laquelle l'activité n'est pas normale. On atteint des résultats très différents dans les recensements agricoles d'après l'époque à laquelle on fixe le dénombrement. Dans certains pays, dans certains états sociaux, la fixation du moment de l'observation peut présenter de grandes difficultés. Ainsi, dans l'Inde, la date du recensement se trouve conditionnée d'une manière très spéciale : il faut choisir une période de pleine lune, afin que les agents recenseurs, au nombre de deux millions, puissent accomplir leur travail durant la nuit; de plus, la date choisie ne doit coïncider avec aucune fête ou foire, ni avec les jours regardés comme favorables aux mariages et aux ablutions dans les fleuves sacrés, ni avec l'époque des pèlerinages vers des sanctuaires vénérés (1).

78. Nous avons dit aussi que le relevé devait envisager la durée des phénomènes; c'est le dernier aspect sous lequel la notion de temps intervienne pour la fixation des caractères du relevé.

Très souvent, on se borne à enregistrer les faits au moment où ils se produisent, sans plus, mais dans d'autres cas, non moins nombreux, la durée pendant laquelle les phénomènes continuent à se manifester est une donnée essentielle à recueillir. Il ne s'agit pas de suivre les faits individuellement en assistant, pour ainsi dire, à leur évolution complète, du moins cette condition n'est pas toujours exigée, ni toujours réalisable. Nous passerons en revue

(1) *Census of India, 1911.* Introduction.

quelques-uns des faits qui donnent lieu à l'observation statistique sous cette forme.

a) Phénomènes dont la durée est relevée par induction. Ce sont ceux dont on ne peut suivre individuellement le développement complet. En premier lieu apparaissent les faits de la vie humaine : la durée de l'existence, des unions conjugales, d'une génération, etc... Il est essentiel, pour la connaissance de ces éléments si importants, d'établir le temps pendant lequel ils existent. Dans ce but, il ne faut pas nécessairement rechercher la durée de l'existence de tous les hommes dont la naissance a été comptée dans les statistiques de la natalité, ni de tous les mariages qui ont été célébrés. Il suffit de procéder par induction et de relever, à une date donnée, celle d'un recensement, l'âge de toutes les personnes recensées. La distribution des unités recensées, par catégorie d'âge, fournira la réponse à la question posée, parce que, par induction, nous passons des faits connus aux faits inconnus en établissant la loi générale du phénomène. Ainsi, une table de survie se construit en se basant sur le tableau de la population par âge : celui-ci est l'image, comme dit Levasseur, d'une population qui comprend une série de générations ayant eu des chances diverses de mortalité et ayant d'ordinaire subi des modifications par l'émigration et l'immigration. La méthode démographique de construction d'une table de survie est basée sur le procédé inductif. Etant connus la population par âges et le nombre des décès par âges, on détermine, pour chaque période, le rapport de ces deux nombres; partant ensuite de la population ramenée à 100,000 personnes, le statisticien déduit les décès afférents à chaque âge jusqu'au moment où la différence entre les deux termes devient nulle (1).

b) Phénomènes dont la durée est relevée par l'observation directe. Un grand nombre de faits rentrent dans cette

---

(1) LEVASSEUR. *La population française*, t. II, pp. 286-290.

catégorie; par exemple, un des éléments servant à caractériser une grève est sa prolongation; dans la statistique des grèves, on relève donc pour chaque conflit le nombre de jours de chômage. Pour les accidents de travail, une donnée essentielle est la durée de l'incapacité, parce qu'elle fait varier l'étendue de la charge; tout accident du travail est porté dans les statistiques avec les indications relatives à cet objet. Les divisions adoptées sont en concordance avec la loi de réparation des accidents. Dans ces cas et dans une quantité de cas semblables, la durée du phénomène est relevée directement par l'observation, sans recourir au procédé inductif.

III. — Le relevé considéré sous le point de vue de l'espace.

79. A proprement parler, l'espace est sans bornes. Au delà de l'univers sensible nous pouvons concevoir une étendue sans mesure assignable, car après un objet on peut toujours supposer qu'il en existe un autre. L'espace dépasse donc les bornes de l'univers et est inconnaissable en lui-même; la seule réalité sous laquelle il nous apparaisse est l'étendue des corps situés dans l'espace et les relations de leurs distances.

Si telle est la notion philosophique de l'espace, nous pouvons en conclure que les divisions d'après lesquelles nous situons les corps et les phénomènes sont artificielles dans leur essence et que nous pouvons les faire varier selon que le sujet de nos études l'exige. Des géomètres ont pu imaginer une géométrie à quatre dimensions, bien que nos sens ne perçoivent que trois dimensions des corps; il serait aussi légitime pour le statisticien d'adopter des mesures de l'étendue appropriées à ses conceptions propres, sans s'inquiéter si elles sont ou non d'un usage courant. Mais, en pratique, on n'a pas à recourir à des mesures nouvelles, celles existantes suffisent à tous les besoins. On peut tracer parmi elles plusieurs catégories.

a) Divisions politiques et administratives. Elles sont des plus usitées dans un grand nombre de domaines explorés par la statistique. Ce sont tout d'abord l'Etat, l'unité politique par excellence, puis la province, département, comté, quel que soit le nom qu'on lui donne, ensuite des divisions administratives plus réduites : cercle, district, arrondissement, canton, etc., enfin, l'unité administrative primaire, la commune.

b) Divisions économiques. Si les divisions politiques conviennent à une foule de recherches statistiques, il existe des investigations qui réclament une base différente, tant pour l'organisation des recherches que pour la présentation des résultats; lorsque le caractère d'une statistique est nettement économique, on se demande pour quelle raison on continue à lui appliquer des divisions d'ordre politique. Ainsi, dans les statistiques agricoles, il semblerait plus sage d'utiliser la division, existante et non à créer, de régions agricoles, plutôt que de s'en référer à une localisation des faits basée sur une conception étrangère au but de la statistique. Tous les pays présentent de vastes divisions agricoles, par nature de terrains, qui répondent mieux que toute autre à une statistique de la production et de l'élevage. Dans les statistiques industrielles, les divisions sont moins nettes et dès lors moins facilement utilisables. Un bassin industriel n'est pas nettement délimité comme peut l'être une région agricole. Cette formule a été employée pour la grande industrie charbonnière et métallurgique, mais elle n'a pas de sens en ce qui concerne les agglomérations urbaines, si fécondes en industries, ni en ce qui regarde les régions de petite industrie. Nous avons, en Belgique, le bassin de Charleroi, du Borinage, du Centre, de Liége, mais il n'existe pas de dénomination semblable pour le Brabant ni les Flandres. Bien que cette division par bassin soit séduisante en théorie, elle n'est pas susceptible d'être utilisée en pratique dans tous les cas. Il faut se borner à en souhaiter l'usage dans des cas spéciaux.

Enfin, on peut concevoir aussi des subdivisions du territoire, ayant un caractère économique, basées sur des relations démographiques. Telles sont celles qui subdivisent la population entre les centres urbains et les régions rurales, entre les grandes villes, celles de moyenne importance et les petites villes, ou simplement répartissent une population en tenant compte du nombre d'habitants des localités où elle vit. Ces divisions sont importantes dans les recherches économiques et sociales; la démographie les utilise dans le but de déduire, des nombres ainsi groupés, des considérations d'ordre économique.

c) Divisions naturelles. Elles traduisent, sous une forme schématique, l'influence des milieux. Elles sont de mise dans les pays présentant une grande superficie, mais non dans des Etats resserrés dans d'étroites limites. A ce point de vue, on peut distinguer entre les régions de plaines et de montagnes, entre les côtes et l'intérieur, entre le Nord et le Midi, entre l'Est et l'Ouest, entre une des régions quelconques et le centre, entre la campagne et la forêt, entre les bassins fluviaux.

On aperçoit immédiatement la richesse des aperçus auxquels on peut se livrer en faisant ainsi varier le point de vue auquel on se place : aussi les divisions de l'espace intéressent-elles la statistique non seulement dans la phase de la réunion des données, mais aussi dans celles de la présentation et de l'interprétation. Dans un pareil système, l'unité d'observation serait la commune, puis on grouperait par régions les communes présentant des caractères analogues.

80. Quelle est la relation systématique existant entre le relevé et les divisions conventionnelles ou naturelles dont il vient d'être question? Nous l'avons exprimée en disant que le relevé doit être aussi étendu et aussi particularisé que possible. Aussi étendu qu'il se peut, en ce sens que les phénomènes à relever doivent l'être partout où ils

se produisent; en les observant partout, on est certain de n'omettre aucune des qualités qu'ils présentent et d'obtenir une mesure exacte de leur fréquence. Aussi particularisé qu'il est possible, afin que les détails s'aperçoivent; qu'ils ne se fondent pas dans une masse peu homogène, sans précision et sans originalité. Les bornes politiques de l'Etat limitent le relevé par en haut, comme l'étendue de la commune ou de la plus petite division administrative le limite par en bas. Les opérations statistiques de grande envergure s'arrêtent à l'Etat. Pratiquement on n'a jamais été plus loin. Ce que l'on désigne du nom de statistiques internationales ne mérite pas ce nom dans l'acception où on le prendrait en parlant du relevé. Ce n'est point un relevé unique, décrété par une réunion d'Etats, s'étendant simultanément aux territoires appartenant à plusieurs gouvernements, exécuté d'après des instructions uniformes, publié sur un plan unique. Les statistiques internationales sont simplement des compilations de statistiques nationales, réunies et mises bout à bout, non sans hiatus, et présentant dans leur ensemble toutes les lacunes des statistiques particulières, en y ajoutant les défauts résultant du manque d'uniformité dans les définitions et des différences dans l'exécution et la mise en œuvre. Ces statistiques sont internationales de nom, nationales de fait. Il est donc permis de dire que l'Etat forme actuellement la limite extrême d'extension des investigations statistiques.

Se trouverait-on encore dans les limites du relevé direct si, au lieu de porter sur l'Etat entier, les statistiques s'arrêtaient aux limites d'une province, ou d'un groupe de provinces ou de divisions analogues? On pourrait être tenté de répondre affirmativement en considérant que le relevé direct embrasse toutes les unités comprises dans une définition générale. Or, une définition pourrait limiter à une province unique les phénomènes à observer. La raison de décider la négative est que, dans un cas semblable, la limitation de l'observation à une seule province quand

le phénomène peut être relevé dans plusieurs ou dans toutes, est arbitraire et serait contraire à la règle suivant laquelle le relevé doit être aussi étendu que possible. On quitterait le domaine du relevé direct pour entrer dans celui de la monographie.

La limite inférieure de l'observation statistique est la plus petite division administrative, la commune. Les divisions économiques et naturelles sont toujours plus étendues que la commune. C'est la vie économique la plus étroitement localisée qui fournit parfois le plus de données intéressantes et c'est en l'observant qu'on se rapproche le plus de la réalité et de la vie. Pour l'étude démographique, nous n'apercevons pas dans tous les cas le même intérêt de choisir la commune comme unité d'espace, parce que la plupart des faits relatifs à la population demandent, avant tout, de vastes ensembles où disparaissent les caractères accidentels, mais pour certaines questions, il n'est pas inutile de pousser jusqu'à là le soin du détail. Ainsi en est-il, par exemple, de la densité de la population : l'existence d'une grande ville suffit à augmenter beaucoup le coefficient de la densité de la population dans un canton ou arrondissement administratif. Dans un même arrondissement administratif — celui de Bruxelles — on notait, en 1912, des différences dans la densité de la population allant de 208 à 28,641 habitants par kilomètre carré dans les différents cantons de milice. La moyenne pour l'arrondissement administratif (967), ne donne aucune idée des différences énormes qui se remarquent dans les parties dont elle se compose.

IV. — Les procédés et les organes du relevé.

81. Pour introduire de l'ordre dans un exposé nécessairement complexe, nous avons indiqué, dans les paragraphes précédents, quels sont les phénomènes à comprendre dans le relevé et les conditions auxquelles ils doivent satisfaire

pour pouvoir être isolés de ceux qui restent en dehors de l'opération statistique. Les limites de temps et d'espace ont aussi été analysées dans leurs caractères généraux. Ce qui précède pourrait constituer la préparation théorique du relevé; la phase exposée dans ce paragraphe peut être considérée comme la préparation pratique. Elle répond à la double question que voici : de quelle nature sont les documents à utiliser pour le relevé? Par quels moyens arrive-t-on à connaître les faits compris dans la sphère des observations? Elle embrasse l'examen de deux groupes de problèmes : les moyens matériels (bulletin, questionnaire, etc.) et les moyens personnels (agents du relevé).

A. — *Le bulletin ou questionnaire.*

82. Tout relevé statistique ne nécessite pas l'emploi ni d'un bulletin ni d'un questionnaire. La distinction établie plus haut entre le relevé automatique et le relevé réfléchi reprend ici toute son importance. Dans le relevé continu, la nature des faits à observer est déjà indiquée par les dispositions législatives ou administratives en vertu desquelles le relevé s'exécute; presque toujours, ces mêmes dispositions prescrivent la façon dont les faits doivent être annotés, consignés dans des registres ou des formules dont la forme est arrêtée d'une manière fixe. Le statisticien n'a pas à s'en occuper directement; il les utilise telles qu'il les trouve. Bien entendu, si, à l'expérience, ces documents sont jugés peu ou mal utilisables par la statistique, on pourra les modifier de manière à donner satisfaction à des réclamations légitimes. Mais ceci est l'exception. En thèse générale, le statisticien est forcé, dans ce domaine, de se contenter de ce qu'il trouve. Il serait désirable qu'un accord s'établît toujours entre les organes administratifs et scientifiques, afin que le fond et la forme des documents du relevé automatique répondent à toutes les exigences scientifiques et notamment que le point de vue mathématique ne soit pas éli-

miné des considérations se rapportant à cet objet. C'est une cause de graves mécomptes et l'occasion de pertes de temps et d'argent inappréciables.

Conçoit-on, par exemple, que l'indication de charges financières à mettre en relation avec le nombre probable de bénéficiaires et l'âge de ceux-ci puisse être arrêtée dans sa forme sans des études préalables à confier à un actuaire? La chose peut paraître inimaginable, mais il en a été parfois ainsi. D'informes tableaux, décorés pour la circonstance du nom de « statistiques », continuent à être remplis avec application par des légions de commis, sans que la science ou même une pratique éclairée puissent espérer en retirer jamais le moindre avantage.

Parmi les documents, dressés dans un autre but, utilisés par la statistique néanmoins, citons les registres de l'état civil, les états dressés en matière de justice criminelle, les documents tenus par les agents du service des douanes et des accises, les registres de population là où ils existent et sont tenus avec soin. On peut aussi se baser sur ces documents, non pour les utiliser exclusivement, mais pour y trouver le point de départ de recherches plus étendues : tel fut le cas, en Belgique, lors du recensement de l'industrie et des métiers au mois d'octobre 1896, où l'on se servit des registres de population pour connaître — en l'absence d'un recensement général des professions — le nom, l'adresse et l'industrie exercée par les patrons comme par les ouvriers; un bulletin détaillé était ensuite remis à ces personnes.

Tous les documents officiels ne méritent pas la même confiance, et dans un même document il se trouve des données méritant moins de créance que d'autres. Il appartient à la critique, dont les principes seront exposés plus loin, d'opérer la ventilation nécessaire entre les différentes données. L'origine officielle du document ne doit pas aveugler le statisticien. D'ailleurs dans un même registre se trouvent confondues des données dont l'origine

est très différente : les unes pourraient être certifiées par le fonctionnaire qui les a recueillies, les autres sont simplement reçues à titre de déclarations faites par les intéressés, sans aucun contrôle par le fonctionnaire qui les reçoit. On en trouve des exemples dans la statistique commerciale où il y a quantité de données dont les unes réunissent des garanties presque absolues, tandis que d'autres présentent un aléa excessif.

83. Pour le relevé réfléchi, organisé dans un but déterminé, l'instrument de l'observation doit être fabriqué de toutes pièces : c'est le *bulletin* ou *questionnaire*. On donne ce nom au document destiné à recevoir les réponses des personnes comprises dans le relevé. Le comptage des unités et l'addition des données numériques contenues dans les réponses données au bulletin permettent de former les masses statistiques qui figurent dans les tableaux de présentation.

Le mot *bulletin* est d'habitude employé pour désigner le document dont on fait usage dans les recherches de grande envergure faisant partie du relevé direct. On dit un bulletin de recensement, non un questionnaire de recensement... Le mot *questionnaire* a un sens plus étendu, il se réfère à un document détaillé, comme on peut en employer dans un relevé s'adressant à un nombre restreint de personnes ; on dit, de préférence, un questionnaire d'enquête. Toutefois la distinction ci-dessus, basée sur l'usage, n'a rien d'absolu.

Le bulletin employé dans les relevés étendus est de deux sortes : le bulletin collectif et le bulletin individuel.

On appelle *bulletin collectif* le document destiné à recevoir des réponses se rapportant à plusieurs personnes formant une collectivité familiale, économique, ou simplement constituant une réunion de fait.

Dans un recensement de la population on peut employer le bulletin collectif, remis au chef de la famille, afin que celui-ci y porte toutes les indications relatives aux personnes

vivant avec lui. Au lieu de la famille, on peut adopter, comme mesure collective, le *ménage,* qui n'implique pas que les personnes vivant ensemble soient unies par des liens de parenté. En Belgique le document du recensement de la population est un bulletin de ménage dans lequel toutes les personnes faisant partie du ménage et ayant leur résidence habituelle dans la maison, y compris celles qui sont momentanément absentes, doivent avoir consigné leur nom de famille, leur prénom, le degré de parenté avec le chef de ménage, leur état civil, profession, fonction ou situation, etc. Il y a, en plus, un bulletin spécial personnel utilisé pour l'inscription des personnes qui, sans avoir leur résidence habituelle dans la maison, s'y trouvent accidentellement au moment du recensement, et un bulletin spécial collectif destiné au relevé des personnes réunies dans les pensionnats, casernes, établissements charitables, etc.

L'Italie avait adopté pour son recensement, en 1881, le bulletin de famille, principalement parce que le bulletin individuel propre à chaque membre de la famille se serait adressé à un trop grand nombre de personnes ne sachant ni lire ni écrire. En 1901 ce pays a substitué le bulletin individuel au bulletin de famille.

Le bulletin individuel ne doit recevoir que les indications relatives à une seule personne. Il est en usage dans presque tous les pays, mais il est presque toujours accompagné d'autres bulletins ayant le même caractère que le bulletin collectif. En France, bien que le bulletin individuel soit l'instrument principal du recensement, il y a aussi un bulletin par famille et un bulletin par maison. De plus, certaines catégories de population doivent être comptées à part, comme les lycées, les séminaires, les hospices, les asiles d'aliénés, etc.

84. Le bulletin collectif ou de famille a été adopté, au début de l'ère des recensements, dans presque tous les pays et il est encore usité dans quelques-uns de ceux qui sont

parmi les plus avancés en fait d'organisation statistique. Dans un état social qui ne comporte pas un développement général de l'instruction, il s'impose afin de ne pas multiplier inutilement les difficultés : dans la famille on a chance de rencontrer quelqu'un sachant écrire et pouvant répondre aux questions du bulletin de ménage, ou l'agent recenseur peut écrire lui-même les réponses qui lui sont données verbalement. Dans l'Inde anglaise le bulletin collectif est l'instrument du recensement. M. Gait, commissaire pour le dernier Census, a exposé avec détail les raisons qui, dans l'Inde ont fait préférer le bulletin collectif au bulletin individuel; il indique, outre les raisons précédentes, l'utilité qu'il y a, au cours de la revision, à comparer entre elles les réponses données par les différentes personnes composant un même ménage; il y ajoute une raison d'ordre matériel, à savoir que les bulletins individuels auraient formé un volume beaucoup plus fort que les bulletins de ménage.

Lors du recensement de 1881, l'Italie avait adopté le bulletin collectif à raison du manque général d'instruction et aussi pour encourager les communes à établir des registres de population basés sur les bulletins collectifs. Cette attente a été complètement déçue, alors qu'en Belgique les registres de population sont tenus dans toutes les communes avec le plus grand soin; pour renouveler les registres de population, il faut nécessairement se baser sur le recensement effectué au moyen du bulletin collectif de ménage.

Le bulletin collectif a beaucoup perdu de sa faveur à raison de ce fait que, dans la suite des opérations statistiques il ne peut être utilisé directement sans de nombreux inconvénients. Aussi, dans les pays qui ont conservé le bulletin collectif le complète-t-on en faisant transcrire sur des fiches les indications relatives à chaque personne, individuellement, comprise dans le bulletin de ménage. Il en a été ainsi aux Indes, où, lors du dernier recensement, on a dû confectionner plus de 315 millions de fiches indivi-

duelles. Depuis 1876, on a renoncé, en Belgique, à procéder au dépouillement des bulletins de ménage par voie de pointage et l'on a chargé les agents recenseurs de reporter sur des fiches individuelles toutes les indications relatives aux personnes figurant sur le bulletin collectif. C'est par le comptage des fiches qu'on détermine actuellement le nombre de personnes de chaque catégorie. Evidemment, la transcription des données sur fiches exige du temps et de l'argent, mais ces inconvénients sont compensés par la sûreté plus grande des résultats : en combinant, par exemple, les couleurs et les profils des fiches, on arrive à une précision qu'on ne pourrait atteindre avec le procédé du pointage. La confection des fiches est comprise dans la rémunération des agents recenseurs. La dépense n'est pas très lourde dans des pays comme l'Inde, mais s'il fallait appliquer ce système dans un pays à hauts salaires la dépense s'élèverait à un chiffre inacceptable.

85. Cette dernière considération a engagé les statisticiens à substituer au bulletin collectif le bulletin individuel. En principe, les indications contenues dans ce bulletin sont relatives à un seul individu, de telle façon qu'elles constituent elles-mêmes des fiches, sans que l'on doive procéder à la transcription des données du bulletin collectif.

Théoriquement, l'introduction de la fiche individuelle est une amélioration sur le système suivi auparavant, mais ce genre de bulletin ne peut être employé utilement que dans les pays où le niveau général de l'instruction est suffisamment élevé. Il n'est pas sans présenter d'ailleurs certains inconvénients : ainsi, l'uniformité extérieure des bulletins est une source d'erreurs dans le classement; si les différentes catégories de personnes pouvaient être distinguées au moyen de la couleur du bulletin, on aurait beaucoup moins de classements défectueux, mais il ne peut être question de charger l'agent recenseur de répartir les bulletins de différentes sortes; ce système comporte-

rait, de plus, une grande part d'arbitraire, chacun utilisant le bulletin qui lui aurait été remis plutôt que de réclamer, en cas d'erreur, un bulletin différent, approprié à la situation réelle (1). On ne peut non plus, sur le bulletin originaire, substituer aux réponses elles-mêmes des signes conventionnels qui facilitent la lecture, comme on le fait lorsque les fiches sont directement confectionnées par le service statistique. Enfin, souvent, l'écriture est illisible, les bulletins sont maculés, d'un maniement difficile et désagréable, les noms sont incorrectement orthographiés, etc.

Tous ces inconvénients, qui sont réels, n'empêchent pas le bulletin individuel de l'emporter sur le bulletin collectif au point de vue de la tabulation.

Le bulletin individuel est strictement propre à une personne seule dans les relevés démographiques. Dans les relevés économiques, il faut un peu étendre les questions qu'il comprend; ainsi, à un chef d'entreprise, on pourra demander non seulement les renseignements qui le concernent personnellement, mais encore ceux qui se rapportent à son entreprise : nombre, sexe, qualité des ouvriers et employés, force motrice utilisée, etc. Ces données ne se rapportent pas strictement à la personne même, mais à l'entreprise dirigée par l'auteur du document; elles permettent d'établir un contrôle entre les réponses individuelles données par les ouvriers d'une part et le chef d'entreprise d'autre part.

86. Comme exemple de bulletin individuel, nous reproduisons, ci-après, le modèle de bulletin employé en France, lors du recensement général de la population effectué le 4 mars 1906.

---

(1) Royaume de Belgique. Ministère de l'Industrie et du Travail (Office du Travail). *Recensement de l'industrie et du commerce*, 1910. Introduction, p. XXI, col. II.

BULLETIN INDIVIDUEL

Il doit être établi un bulletin pour chaque personne qui a passé dans la maison la nuit du 3 au 4 mars 1906, même pour les enfants en bas-âge.

1. **Nom, sexe.**	Nom de famille Prénoms — Sexe (m. ou f.)
2. **Date et lieu de naissance**	Né en l'année 1 , le — Indiquer la colonie ou dans la commune d — le pays étranger :
3. **Nationalité**	Etes-vous { né Français? ou naturalisé Français? étranger ? de quel pays?
4. **Etat de famille. . . .**	Etes-vous : célibataire? marié? veuf? divorcé ? Depuis combien d'années avez-vous été marié ? Combien avez-vous eu d'enfants? Encore vivants ? Morts (mort-nés non compris)?
5. **Résidence.**	Habitez-vous dans la commune? Ou y êtes-vous de passage ?
6. **Instruction**	Savez-vous lire et écrire ?
7. **Profession.**	Quelle est votre profession principale? Si vous en exercez une autre, laquelle ? (Indiquer la profession, l'industrie, le commerce, sous une forme précise et détaillée; si vous n'exercez aucune profession, répondre néant et ne pas répondre aux questions suivantes).

8. — Renseignements sur la profession ou l'industrie principale à la date du recensement.

Si vous êtes patron, chef d'établissement, dans une profession agricole, industrielle, commerciale, libérale, ou ouvrier à façon travaillant chez vous, répondre aux questions suivantes :

a) Raison sociale, nom, adresse de l'établissement ou de l'entreprise que vous dirigez :
Rue , n°
Commune d
Arrondissement d

b) Combien de personnes occupez-vous actuellement, au total, dans cet établisseme

(Ne pas comprendre les patrons, ni les ouvriers à domicile, les domestiques exclusivement attachés à la personne).

c) Etes-vous ouvrier à façon travaillant à votre domicile?

Si vous travaillez sous la direction ou au service d'autrui, comme ingénieur, employé, ouvrier, journalier, garçon, apprenti, domestique, etc.

(Les ouvriers à façon à domicile doivent s'inscrire dans la partie de gauche.)

a) Nom et adresse du patron, de l'entreprise, de l'administration qui vous emploie :
Rue
Commune d

b) Nature de la profession, de l'industrie, du commerce de votre patron ou de l'administration qui vous emploie :

c) Si vous êtes sans emploi :
Depuis combien de jours?
Est-ce pour cause de maladie ?

En même temps que le bulletin individuel, principal instrument employé pour le recensement, le dénombrement de la population en France utilise deux bulletins collectifs : une feuille de ménage et un bordereau dressé pour chaque maison. Ce dernier document, en forme de chemise, doit contenir toutes les feuilles de ménage de la maison ; il est dressé par l'agent recenseur d'après les résultats portés sur les feuilles de ménage, après enquête sur place, si c'est nécessaire, pour compléter le tableau. La feuille de ménage doit être remplie par les recensés eux-mêmes. Voici le modèle de la feuille utilisée pour le recensement de 1906 :

Feuille de Ménage

Ménage de M . . . .
Rue. . . . .
Profession . . . .
Nombr de personnes composant le ménage (présentes ou absentes, mais non compris les personnes de passage). . . . .

Habitation. — Nombre de pièces destinées à l'habitation des membres du ménage :

(On comptera comme pièce tout compartiment d'une maison, destiné à l'habitation, séparé des autres par des cloisons allant jusqu'au plafond et pouvant recevoir un lit d'adulte.) Comprendre la cuisine, l'antichambre, les cabinets de toilette assez grands pour pouvoir contenir un lit, les chambres de domestiques, même séparées du logement, mais non les lieux d'aisances; ne comprendre les boutiques, ateliers, etc., que lorsque une ou plusieurs personnes y passent habituellement la nuit.

Liste nominative des membres du ménage :

Les noms doivent être inscrits dans l'ordre suivant : 1° le chef de ménage (père ou mère de famille); 2° la femme; 3° les enfants; 4° les autres parents faisant partie du ménage; 5° les domestiques; 6° les personnes étrangères à la famille.

Numéros	Nom de famille	Prénoms	Age	Relation de parenté ou autre avec le chef de ménage.	Signaler ici les membres du ménage qui seraient aveugles et ceux qui seraient sourds-muets.
		1° Membres du ménage présents :			
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

2° Membres du ménage absents :

On comprendra dans cette section les membres du ménage en voyage ou malades dans les hôpitaux, ou travaillant au dehors, mais on n'y comprendra pas les enfants placés chez une nourrice, les militaires, les élèves internes des établissements d'instruction publics ou privés, les individus en prison ou dans les hospices et asiles d'aliénés.

1					
2					
3					
4					
5					

3° **Hôtes de passage** (voyageurs, militaires en permission, élèves internes en congé, etc.) :

1					
2					
3					
4					
5					

87. La rédaction du bulletin ou questionnaire ne peut être isolée du programme général, ni de la mise en œuvre des matériaux. Le bulletin, par les questions y contenues, traduit la pensée présidant à l'enquête et a pour but de réunir les données à grouper dans les tableaux. Il y a une union intime entre ces trois phases du travail; aussi est-il à souhaiter qu'elles soient l'œuvre de la personne chargée de la direction.

Les qualités essentielles du bulletin ou questionnaire, au point de vue de sa rédaction, sont les suivantes :

1° Le bulletin doit être succinct, il doit se borner à demander le strict nécessaire. Il ne suffit pas qu'une donnée soit intéressante pour qu'elle ait droit à prendre place parmi les questions à poser. Elle doit, de plus, fournir un élément essentiel en vue de l'enquête et cette donnée doit être susceptible de trouver place dans les tableaux statistiques. Si cette dernière condition ne peut être réalisée, il est inutile, en général, de poser la question, sauf s'il s'agit d'une question de contrôle. L'art ne consiste pas à multiplier les points de vue, mais à faire concorder toute chose vers un but unique, en tenant compte du caractère numérique des données. Les éléments trop difficiles à réunir ne doivent pas être compris au nombre des questions à poser. Il ne servirait à rien de les obtenir en un endroit si on doit y renoncer à un autre, de les faire admettre par les uns si les autres se refusent à y répondre. La brièveté est une des conditions du succès;

2° Il doit être clair, de façon à pouvoir être compris de tout le monde. Il ne faut pas perdre de vue qu'un bulletin statistique s'adresse à des millions de personnes, dont le plus grand nombre ne possèdent qu'une instruction rudimentaire. Si les questions sont conçues en termes trop élevés, elles risquent de n'être pas comprises. Non seulement les recensés, mais aussi les agents recenseurs doivent comprendre exactement la signification des questions du bulletin, de

façon à donner les éclaircissement qui leur sont demandés ou à formuler eux-mêmes la réponse.

L'interprétation donnée à certaines demandes est parfois bizarre; à la question : « Travaillez-vous seul? sinon, combien employez-vous à présent chez vous de membres de votre famille? d'ouvriers salariés? », un bon nombre d'ouvriers à domicile, sans aide ni ouvriers, répondirent qu'ils ne travaillaient pas seuls, mais dans les locaux où d'autres personnes vaquaient à leurs occupations. Lors du recensement belge de 1910 (industrie et commerce), beaucoup d'ouvriers, au lieu de mentionner le nom de leur patron, sa profession et son adresse, comme on le leur demandait, répondirent en donnant le nom de l'administrateur délégué, du directeur-gérant ou même de l'ingénieur dirigeant l'entreprise ou la division à laquelle ils étaient attachés. D'autres, au lieu de mentionner la localité où ils allaient travailler, indiquèrent le domicile de leur patron, habitant une commune autre que celle où l'usine se trouvait établie;

3° Les demandes ne doivent pas pouvoir être interprétées de plusieurs façons;

4° Il sera simple et catégorique, en ce sens, qu'il ne comportera pas de réponses vagues, mais autant que possible des données chiffrées, ou des réponses par *oui* et par *non* ne se prêtant pas à des appréciations subjectives variables. L'idéal serait que la demande soit assez claire pour que chacun fût dispensé de recourir aux commentaires et aux instructions:

5° On n'oubliera pas, enfin, que tous les éléments du problème à résoudre doivent se trouver réunis dans le questionnaire. La statistique ne peut rien donner de plus que ce qu'on en a recueilli à son intention; comme le dit Pidgin, « le fleuve ne peut pas couler plus haut que sa source ».

88. Au nombre des qualités du bulletin, on peut aussi compter la bonne disposition typographique des questions

posées. Le format du bulletin ne peut être ni trop grand, ni trop petit. Trop grand, le bulletin est difficile à manier, trop petit il est incommode à classer. Si l'on veut faire servir le bulletin comme fiche, lors du dépouillement, on doit éviter qu'il contienne plusieurs pages ; si on ne peut se soustraire à cet inconvénient, il serait bon de disposer à gauche, vers le centre du cahier, une ligne de perforation, de façon à permettre de l'utiliser feuille par feuille en guise de fiche. Le meilleur bulletin est celui qui tient sur une page, imprimée d'un seul côté. Si l'on doit s'adresser aux recensés en deux langues, chaque côté de la feuille peut recevoir un texte différent.

La disposition typographique doit être surveillée très attentivement. En regard de chaque question il faut laisser une place suffisante pour permettre l'inscription des réponses par des gens peu habitués à manier la plume. Le questionnaire doit être ligné. Les questions de même nature seront rapprochées les unes des autres. Une séparation bien nette sera marquée entre les diverses questions. On utilisera, pour attirer l'attention, les caractères gras et italiques.

B. — *Agents du relevé.*

89. Avec le travail de rédaction du bulletin, comme avec celui de la préparation du plan et de la détermination des unités de lieu et d'espace, nous sommes déjà sortis du domaine de la théorie. Le choix des agents, les qualités qu'ils doivent réunir, les instructions à rédiger à leur usage, dépendent entièrement du domaine de l'expérience et de la pratique. Il ne faut pas croire que l'aspect théorique des problèmes fondamentaux de la statistique soit beaucoup plus abstrait que leur face pratique ; tout au contraire, un statisticien vraiment averti des difficultés de sa branche, sera souvent d'avis que les problèmes les plus délicats sont ceux où il faut tenir compte de l'équation personnelle de l'enquêteur et du recensé.

Le bulletin est l'instrument matériel de la statistique; l'agent recenseur en est l'instrument personnel. Les deux sont habituellement interdépendants, mais quelquefois, on peut faire parvenir aux intéressés les documents statistiques qu'ils ont à remplir sans avoir recours à l'intermédiaire d'un agent. Ceci nous amène naturellement à considérer les méthodes générales adoptées pour la distribution des bulletins ou questionnaires.

Pour cet exposé, il nous faut un peu dépasser notre cadre et dire quelques mots de certaines méthodes dont nous aurions dû parler à propos du relevé indirect. Une méthode fort simple, dont on a fait un grand usage aux Etats-Unis, consiste à envoyer aux personnes les plus qualifiées des questionnaires qu'elles sont priées de remplir : l'envoi et le retour des questionnaires se font par la poste, ce qui supprime toute intervention d'un agent enquêteur. Le choix des personnes à qui l'on s'adresse n'est pas aussi difficile qu'on pourrait le supposer, pourvu qu'on puisse se guider sur un recensement général assez récent. D'après le recensement belge de 1910, il suffit de réunir trente sous-groupes de la nomenclature sur 71 pour former un ensemble réunissant plus de 82 p. c. du personnel total et de la population ouvrière; en s'adressant aux industriels les plus notables appartenant à ces industries on réunirait facilement des données intéressant une très importante fraction de la population industrielle. Mais cette méthode ne peut s'appliquer qu'à des enquêtes. En tant qu'elle porte sur des têtes choisies, elle n'appartient pas au relevé direct, lequel exige des investigations s'étendant à toutes les unités, si minimes qu'elles soient. Elle présente encore un autre désavantage : c'est l'énorme déchet auquel elle donne lieu (1).

---

(1) En 1879, pour une enquête intitulée « Testimony of workingmen », le Bureau de statistique de Boston (Mass. U. S. A.) envoya six mille questionnaires à des ouvriers, mais l'enquête échoua tant il y eut peu de réponses. Une enquête sur le travail dans les prisons, organisée sur la même base,

N'étant aucunement astreints à donner une réponse, les industriels en grand nombre s'abstiennent, par négligence ou pour tout autre motif, de faire parvenir leurs réponses au département de la statistique. Enfin, les réponses, plus ou moins rares, qui parviennent aux organisateurs de l'enquête sont loin d'offrir des garanties suffisantes d'exactitude et de sincérité.

On peut recourir à un autre système : s'adresser à des personnes choisies, mais au lieu de leur faire parvenir des questionnaires, les faire visiter par des enquêteurs appartenant à l'administration ou simplement délégués par celle-ci (méthode des agents spéciaux). Ce procédé donne d'excellents résultats, mais l'application qu'on peut en faire est fort limitée : d'abord à raison des frais de voyage élevés qu'elle entraîne, ensuite à cause des délais fort longs qu'elle impose; les agents spéciaux d'ordinaire sont fort peu nombreux et il faut se résigner à laisser s'écouler de longs mois avant d'avoir réuni le matériel statistique (1).

Il est facile de voir que ces méthodes s'appliquent à des enquêtes ou relevés partiels. Par leur nature, elles ne peuvent convenir au relevé direct visant toutes les unités existantes à un même moment dans toute l'étendue d'un pays. Dans une contrée aussi exiguë que la Belgique, on compte 250,000 entreprises industrielles en activité : quel moyen d'obtenir 250,000 réponses volontaires à des questionnaires transmis par la poste, ou de faire visiter à domicile un pareil nombre de chefs d'entreprise?

90. Le relevé direct exige la nomination de véritables agents recenseurs chargés de recueillir les réponses de tous

---

donna quelque résultat après seulement que le Bureau de statistique eût expédié plus de trois cents lettres de rappel. (PIDGIN, *Practical statistics*, p. 7. Boston, 1888.)

(1) Nous sommes entrés dans des détails plus étendus dans notre communication à l'Institut international de statistique sur la « méthodologie générale de la statistique du travail » (*Bulletin de l'Institut international de statistique*, t. 14, 4e livraison, 1905).

les intéressés ou de consigner dans les formulaires de l'enquête les données dont ils auraient connaissance à raison de leurs fonctions. Les lignes qui précèdent indiquent sommairement une distinction essentielle entre le relevé automatique et le relevé réfléchi.

Dans le relevé automatique, les faits sont portés à la connaissance des autorités, sans qu'aucune démarche spéciale soit nécessaire à cet effet. Les statistiques des naissances, des décès, des mariages, des divorces sont constituées par la réunion des données inscrites aux registres de l'état civil; il ne faut donc pas nommer d'agents recenseurs pour s'enquérir des faits, mais seulement organiser la transcription des documents dressés par l'officier de l'état civil. En Belgique, cette transcription se fait par les soins des administrations communales en utilisant des tableaux dressés par le département ministériel ayant la statistique générale dans ses attributions (1). En tant que chargés de la transcription des données qui sont ensuite travaillées par l'administration centrale, les agents des administrations communales remplissent une fonction qui appartient d'ordinaire aux agents recenseurs, mais à cela — et c'est un point accessoire — se borne l'assimilation qu'on peut établir entre eux.

Il n'y a de véritables agents recenseurs que dans le relevé direct.

Ce relevé peut s'étendre à toute la population, comme il peut se restreindre à toutes les manifestations d'un seul fait économique, politique, financier. Plus les faits sont complexes, plus leurs manifestations sont nombreuses, plus aussi augmente la difficulté de recruter des agents capables et attentifs. Moins favorisé que le savant qui, dans son laboratoire, a entre lui et l'objet étudié un seul intermédiaire, le microscope, le statisticien professionnel ne voit rien,

---

(1) « Mouvement de l'état civil et de la population en Belgique pendant les années 1876 à 1900 », par C. JACQUART. *Bulletin de la Commission Centrale de statistique,* t. XIX, p. 298. Bruxelles, Hayez, 1906.

n'observe rien par lui-même : des milliers d'yeux voient pour lui, des milliers d'intelligences interprètent sa pensée, mais ce n'est pas — bien au contraire — un avantage ! Pour remédier au manque d'unité, il faudra prendre de multiples précautions. Nous en avons déjà indiqué quelques-unes en parlant du bulletin : concision, simplicité, caractère précis des questions. Nous résumons ci-après celles qui sont de mise à l'égard des agents.

91. Selon la formule donnée par Gabaglio, les agents du relevé réuniront trois conditions : A, *savoir,* B, *pouvoir,* C, *vouloir;* le tout s'appliquant aux données statistiques dont il s'agit de réunir les éléments.

A. — Dans la première condition, nous pouvons distinguer des qualités propres à l'agent, et d'autres acquises à la suite des instructions rédigées par la direction du dénombrement. L'agent recenseur doit posséder une instruction générale suffisante ; des qualités d'ordre, de conscience et de soin sont indispensables pour l'accomplissement de sa tâche. A une connaissance parfaite de la langue parlée par les recensés, — même des patois locaux — il joindra, si c'est possible, une connaissance non moins étendue de l'endroit où il a à remplir sa mission. La désignation, en qualité d'agents recenseurs, de petits fonctionnaires de l'Etat ou des communes, donne déjà des garanties au sujet de leur instruction ; leur nomination par les autorités communales, sous le contrôle du gouvernement, est prescrite en vue du choix de personnes connaissant parfaitement la localité. Il est recommandé de choisir les agents recenseurs parmi les fonctionnaires communaux, les instituteurs, les agents pensionnés. Le recensement belge de l'industrie et du commerce de 1910, prescrit aux administrations communales de désigner, pour effectuer le recensement, des agents instruits et capables, choisis autant que possible parmi les secrétaires communaux, les instituteurs, les fonctionnaires retraités ou d'autres personnes exerçant ou ayant exercé

des fonctions propres à faciliter l'accomplissement de leur mission (arrêté royal du 15 décembre 1910, art. 17). Dans les grandes villes, à cause du nombre d'agents recenseurs à nommer, ces fonctions sont surtout confiées à des agents de la police. On a émis le regret, en France, que dans les grandes villes, le nombre des agents recenseurs soit tout à fait insuffisant (1). On a aussi tenté de recruter un corps volontaire de recenseurs, à qui des récompenses honorifiques seraient accordées; c'est le cas en Allemagne, mais la séduction qu'exercent les emplois officiels n'est pas la même dans tous les pays. S'il est permis d'exprimer une opinion assez paradoxale au sujet de la capacité intellectuelle à exiger des agents recenseurs, nous voudrions dire que les plus intelligents ne sont pas les meilleurs. Ce sont les plus soigneux et les plus attentifs qui répondent le mieux à ce qu'on en attend. Un homme très intelligent jugera peut-être inutile de prendre connaissance avec soin des diverses instructions et il sera tenté de substituer ses vues propres à celles de la direction; on devra compter avec lui sur une équation personnelle très marquée. A l'occasion du dernier recensement aux Indes, on a observé que les bulletins remplis par les simples agents recenseurs indigènes étaient dressés avec plus de soin que les réponses données par les Européens. Pour ceux-ci, le recensement était affaire de mince importance et toute réponse était bonne pour en finir avec une formalité aussi ennuyeuse; l'agent recenseur indigène, au contraire, exerçant une fonction publique, l'a remplie en conscience, se conformant en tous points aux directives reçues.

Il appartient à la direction du recensement de donner des instructions suffisantes et surtout très claires aux agents recenseurs. Ce point mérite quelque développement et pour ne pas scinder ce qui a trait aux qualités personnelles des agents, nous en parlerons plus loin, au n° 92.

---

(1) Compte rendu du recensement de 1906, p. 3, note.

B. — L'agent doit pouvoir exécuter ce qui lui est commandé. Ceci suppose une sage discrétion du questionnaire. Ne faites pas poser par le recenseur des questions indiscrètes, ou heurtant les intérêts ou la conscience si vous voulez qu'il réussisse dans sa mission. Sa tâche, non plus, ne doit pas être trop lourde. Tous les recensements limitent le nombre de bulletins à confier à un seul agent recenseur. Le maire, en France, doit désigner des recenseurs à raison de 1 par 100 habitants dans les campagnes, et 1 par 200 habitants dans les villes; les agents contrôleurs sont au nombre de 1 par 2,000 habitants.

C. — L'agent doit vouloir arriver à un résultat exact. On ne lui laissera pas émettre d'appréciations subjectives, de crainte des divergences de vues et des confusions. La négligence ou l'indifférence des agents est un écueil fréquent; l'organisation d'un contrôle minutieux avec la sanction du refus de l'indemnité si le travail est mal exécuté constitue un moyen d'action efficace. Il faut rémunérer équitablement les agents recenseurs, si l'on veut éviter le dégoût et l'indifférence. Dans certains pays, cette rémunération est à charge des communes; ailleurs, elle est supportée par le gouvernement, comme tous les autres frais du recensement.

92. Dans les instructions aux agents, il ne s'agit plus de conceptions abstraites, mais de réalités. Nous plaçant résolument devant les faits, nous nous demanderons quelles difficultés rencontreront les agents recenseurs, s'ils trouveront dans les instructions les moyens de les surmonter et de quelle façon ils comprendront les règles tracées à leur usage. S'il nous est permis d'invoquer à ce propos notre expérience personnelle, nous dirons que dans les nombreux cas où nous avons eu à rédiger des instructions pour les agents d'une enquête, nous avons toujours procédé de la sorte : *a*) quelle est la nature des difficultés de fait que l'agent est exposé à rencontrer? *b*) les difficultés qui nais-

sent de cette situation de fait sont-elles expressément prévues? *c*) étant donné ce qui est écrit dans les instructions, comment pourrait-on s'y prendre pour le mal comprendre? — car, il ne suffit pas aux instructions d'être claires, elles doivent ne pas pouvoir être détournées de leur sens. On trouve des gens doués d'un véritable génie d'incompréhension ou de détournement du sens naturel des mots. Tous les malentendus ne peuvent être évités; raison de plus pour le rédacteur des instructions de faire tout le possible pour en diminuer le nombre; l'épreuve, que nous appellerons négative, des instructions consistant à s'efforcer de les interpréter à rebours, est un exercice excellent; si les instructions y résistent, elles sont bonnes.

Pour éviter autant que possible les erreurs, divers procédés ont été recommandés et adoptés. Les commissions locales et provinciales de statistique, qui, à l'imitation de la Belgique, ont été créées dans plusieurs pays, ont cessé de fonctionner ou ont été expressément abolies pour des raisons qui se sont imposées partout : la difficulté de trouver des personnes compétentes, dans les petites communes; le fait que, pour toutes les recherches spéciales, on aura plus d'avantage à consulter des fonctionnaires attachés au service en cause, plutôt qu'une commission composée de membres ayant tous la même compétence générale. Actuellement, on s'efforce de faire pénétrer les instructions dans tous les rangs des agents recenseurs en nommant des agents plus qualifiés, en moins grand nombre, dont la mission est de guider les agents recenseurs et de surveiller leur travail. Le bureau central organise aussi des conférences où les différents points du questionnaire et des instructions sont analysés et expliqués. Les conférences donnent de bons résultats, car une des grandes difficultés consiste à obtenir que les agents recenseurs lisent attentivement leurs instructions; le commentaire parlé supplée à la lecture inattentive.

Les instructions à donner aux agents recenseurs comportent deux genres de renseignements : 1° l'interprétation

des termes utilisés dans le bulletin; 2° l'ordre des opérations et la date à laquelle chacune doit être commencée et achevée.

Voyons quelles sont, en France, les opérations confiées aux agents recenseurs. Leur première besogne consiste à établir un document provisoire appelé « carnet de prévision » servant à préparer le travail de recensement et à faire connaître le nombre de bulletins à distribuer. Quinze jours environ avant la date fixée, l'agent recenseur doit aller de maison en maison, s'assurer si l'immeuble est habité ou non, indiquer le nombre de ménages qui y logent et le nombre de membres du ménage. La profession du chef de ménage doit être indiquée d'une façon très précise. Les instructions spécifient ce qu'il faut entendre par maison, ménage, membres et chef du ménage. Cette besogne préparatoire doit être terminée au plus tard le 19 mars.

Vient ensuite la remise des bulletins individuels imprimés; elle s'effectue à partir du 22 mars et l'agent doit déposer dans chaque maison un nombre de bulletins dépassant d'environ un quart celui qui était prévu sur le carnet, afin de tenir compte des hôtes de passage. Les instructions prévoient ce que l'agent a à faire quand il s'agit d'hôtels garnis, de maisons où il existe un concierge, d'établissements dont la population est comptée à part (lycées, casernes), etc...

La reprise des bulletins commence deux jours plus tard. Le principe essentiel à relever ici, c'est que l'agent doit s'adresser personnellement aux intéressés pour recueillir lui-même les renseignements manquant sur les bulletins et rectifier les réponses insuffisantes ou inexactes. L'agent recenseur n'est pas un facteur des postes chargé de distribuer machinalement des imprimés et de les reprendre ensuite sans exercer aucun contrôle. L'agent doit signer le bulletin; c'est un moyen de contrôle. Malheureusement, il faut bien avouer que ce contrôle personnel par l'agent re-

censeur est sans doute l'une des formalités les plus imparfaitement exécutées qui soient. Si les agents recenseurs appliquaient à la lettre les instructions qui leur sont données sur ce point, la qualité des recensements se trouverait du coup améliorée dans une large mesure.

Les délais impartis aux agents recenseurs pour la reprise des bulletins sont de trois jours, en principe, mais on peut aller jusqu'au huitième jour, dernier délai, en cas de difficulté exceptionnelle. Ces délais assurent, dans une certaine mesure, le synchronisme des données du recensement, mais pour être vraiment efficaces elles devraient être plus rigoureuses. Nous avons déjà dit que le recensement hindou, en dépit de toutes les difficultés, est un modèle de recensement synchronique.

SECTION II

Le relevé indirect

CHAPITRE PREMIER

Généralités, définition, divisions

I. — Le relevé indirect et l'induction.

93. Les auteurs ont appelé relevé indirect celui dans lequel toutes les unités ne sont pas comprises, mais d'après les résultats duquel on tente, par induction, de parvenir à la connaissance de la totalité des faits. Si nous connaissons exactement la population de la dixième partie d'une région dans laquelle la densité de la population est constante et égale, nous pouvons nous appuyer sur cette donnée incomplète pour calculer la population totale. Nous ferons tantôt la critique de cette conception; en attendant, nous prions le lecteur de ne pas perdre de vue les conditions de l'hypothèse posée. Le relevé indirect est une simple application des procédés inductifs; il consiste à appliquer au tout, moyennant un procédé de calcul arithmétique, ce qui a été constaté pour la partie à l'aide d'un relevé direct de ses unités ou parties. Du connu, le statisticien remonte à l'inconnu; il suppose exact pour le reste ce qu'il a constaté pour une partie. Entre le connu et l'inconnu, il existe une relation numérique à exprimer au moyen d'un coefficient. On peut admettre un coefficient unique, ou différentiel. Dans les deux cas, le caractère du relevé ne change pas, c'est toujours une application de l'induction.

Depuis Moreau de Jonnès, les traités de statistique ont relaté les calculs de Vauban pour connaître la production agricole de la France, ceux de Necker qui remplaçait un recensement de la population par l'établissement d'un coefficient de natalité, ceux encore de Lavoisier essayant de déterminer l'étendue des terres arables au moyen du nombre de charrues.

Sans nous attarder à reproduire ici ces expériences auxquelles on serait parfois tenté de donner le nom de bizarreries, examinons uniquement les caractères généraux de l'induction et voyons si les conditions du relevé partiel satisfont aux exigences théoriques.

L'induction suppose l'observation de faits positifs, notés avec leurs caractères de fréquence, de grandeur, d'espèce, faits répétés un grand nombre de fois dans des conditions régulières excluant l'intervention de phénomènes accidentels.

La seconde phase du raisonnement inductif est l'hypothèse : la répétition des mêmes faits, ou ce qui revient au même la répartition régulière des espèces d'un seul fait, est l'indice de l'existence d'une loi du phénomène, qu'il est possible de vérifier en dehors du champ actuel d'observation.

A l'hypothèse succède la vérification par observation ou par expérience; enfin, vient la déduction.

On voit très vite que le relevé indirect ne répond que très imparfaitement aux conditions théoriques de l'induction. En premier lieu, nous ne pouvons savoir si les faits observés sur une étroite partie du territoire, ou parmi une infime fraction des unités existantes, appartiennent ou non à la catégorie des faits susceptibles d'une répétition normale. On voudra bien nous épargner l'objection consistant à confondre la nature du fait, qui n'est pas en cause, avec sa mesure statistique, la seule chose dont il s'agit ici. Certes, il y a des décès, des mariages, des naissances partout; il y a partout des salaires relativement élevés et des salaires

minimes. La question n'est pas là; elle est de savoir si la répartition des salaires est la même dans tout le pays que dans telle province, si elle est identique dans les industries non enquêtées et dans celles qui ont fait l'objet d'une étude, si la proportion à la population pour les mariages, les naissances et les décès est la même dans tout le pays que dans telle province observée à l'aide de la statistique. Or, *a priori,* nous pouvons répondre négativement à cette question. Il n'est pas vrai que cette proportion soit applicable telle quelle à d'autres parties que la partie soumise à l'observation. Nous avons vu que la complexité des faits crée des combinaisons innombrables (Cfr. Introduction, ch. III). Comment ces combinaisons pourraient-elles se trouver réalisées, dans un espace étroit, en si grand nombre qu'elles formeraient une image typique du reste? Plus raisonnable apparaît l'hypothèse qu'à chaque condition de milieu, de race, de climat, de culture, d'industrie, d'outillage, de marché, correspond un « jeu » particulier de combinaisons et de causes dont l'ensemble donne au pays, ou à de vastes régions du pays, le caractère qui leur est propre. Lorsque Lavoisier établissait le rapport entre l'étendue cultivée de la France et le nombre de charrues, a-t-il pu tenir compte de la différence des cultures et du sol, des habitudes rurales comme des assolements, de l'étendue du mobilier agricole comme de l'abondance des animaux de trait? Il eût été impossible qu'il le fît, et surtout qu'il réussît à exprimer de telles constatations en un simple coefficient ou rapport.

En second lieu, ce qui manque au relevé indirect pour rentrer dans les conditions du procédé inductif, c'est la vérification, véritable nerf de l'induction. Entre l'observation et la déduction, le statisticien passe parfois trop rapidement sur l'hypothèse en omettant tout à fait la vérification. Le procédé inductif se base sur les caractères d'identité; il manque donc de sûreté quand il s'applique à des faits qui, comme le dit Gabaglio, montrent plus de différences qu'ils n'ont entre eux de ressemblance, qui sont susceptibles d'in-

finies variations comme le sont tous les faits de la vie sociale.

II. — Divisions du relevé indirect.

94. On distingue le relevé indirect par estimation et le relevé indirect proportionnalisé.

Le relevé indirect, surtout par estimation, a été très fréquemment employé avant le XIXe siècle. Un grand nombre de statistiques anciennes sont basées sur de simples évaluations et, comme on nous en avertit rarement, nous sommes parfois tentés d'attribuer à ces chiffres une rigueur qu'ils n'ont pas et qu'ils ne peuvent atteindre. La population des villes au moyen-âge a été souvent établie par des procédés peu sûrs assimilés à tort à ceux en usage à l'époque actuelle.

Dans le relevé par estimation, on procède :

A. — Par approximation. Cette méthode se base sur la connaissance du sujet que l'observateur est supposé posséder. Les indications fournies constituent des renseignements plus ou moins imparfaits, suffisants pour obtenir une vue d'ensemble et une comparaison en général. La statistique agricole utilise fréquemment ce procédé; il peut avoir son utilité quand il s'agit d'obtenir très vite des renseignements qui ne doivent servir qu'à indiquer une tendance.

B. — Par analogie. Ce procédé consiste en une étude soigneuse des conditions numériques d'un certain nombre de faits d'un ordre déterminé et dans la déduction qu'on en tire à l'égard des conditions quantitatives de faits d'un ordre semblable.

Dans le relevé proportionnalisé, on procède :

A. — De la partie au tout. Une partie seule est soumise à l'observation. Les chiffres trouvés pour cette partie sont ensuite appliqués à la totalité. La statistique postale est

basée sur deux comptages faits pendant une durée de sept jours, à deux époques différentes de l'année ; la moyenne de ces deux relevés, multipliée par 52, fournit le chiffre des correspondances pour l'année entière.

B. — D'un phénomène à l'autre. La condition essentielle est qu'il existe une relation étroite entre les deux ordres de faits considérés. De la connaissance précise qu'on a d'une chose, on déduit les caractères d'une autre chose. Ainsi, on a toujours connu facilement le nombre des naissances et des mariages, tandis qu'on a été longtemps sans savoir le chiffre de la population. Lorsqu'on a pu établir un rapport entre ces divers éléments, le chiffre de la population était supposé être celui qu'on obtenait en multipliant le nombre des naissances par le coefficient calculé précédemment.

CHAPITRE II.

L'enquête et la monographie

I. — L'enquête.

95. Les enquêtes sont des recherches offrant un caractère moins général que les travaux statistiques proprement dits. Elles ne cherchent pas à embrasser toutes les unités existantes, à les mesurer et à les dénombrer toutes. C'est à raison de ce caractère que l'enquête appartient au relevé indirect.

On distingue les enquêtes privées et les enquêtes publiques.

Les enquêtes privées rencontrent de nombreux obstacles, car les moyens d'investigation dont les individus isolés disposent sont d'ordinaire fort restreints. Sauf des cas exceptionnels, pour les recherches étendues, il faut nécessairement l'organisation, les ressources et l'autorité de l'Etat.

Les enquêtes publiques sont organisées par les pouvoirs publics ; on peut en distinguer deux types principaux :

1° Les enquêtes parlementaires. Elles ne sont pas exécutées par les assemblées législatives, évidemment inaptes à cette mission, mais par des délégués de ces assemblées, ou simplement à leur demande et d'après un programme tracé par elles. Les enquêtes parlementaires peuvent être libres ou obligatoires; dans ce dernier cas, les témoins cités sont tenus de répondre aux questions posées. L'Angleterre est le pays par excellence des enquêtes parlementaires : les questions soulevées par la réglementation du travail en ont nécessité un grand nombre. Aux époques de crise économique, il y a eu dans plusieurs pays des enquêtes sur la condition des classes ouvrières;

2° Les enquêtes statistiques. Ce sont des travaux statistiques entrepris par des administrations spéciales dans le but de préciser certains phénomènes sociaux, mais sans recourir au dénombrement complet de toutes les unités, ou n'envisageant que des aspects spéciaux de la question.

96. En quoi l'enquête diffère-t-elle du relevé? C'est une question intéressante sur laquelle on a passé peut-être un peu rapidement. Nous essayerons d'en donner un aperçu systématique.

La matière des enquêtes ne semble différer en rien de celle des statistiques. L'enquête concerne les caractères des sociétés humaines, leurs aspects généraux, leurs manifestations typiques, leurs tendances. Elle s'est dirigée, il est vrai, plutôt vers les problèmes économiques que vers les questions démographiques, mais il n'y a pas là matière à une véritable différenciation. Il est tout naturel que l'enquête ait été l'instrument de recherches préféré dans les investigations sur les heures de travail, les salaires, les budgets ouvriers, car il fallait pouvoir multiplier les coups de sonde et l'on ne pouvait guère attendre l'arrivée des grands recensements; ceux-ci, d'ailleurs, avaient leur sphère d'action nettement circonscrite dont il ne pouvait être question

de les faire dévier. En tant que matière, statistiques et enquêtes sont donc des équivalents.

Sous le rapport de la forme, nous notons une première différence, mais les avis seront partagés sur son importance. La statistique traduit en nombre tous ses résultats; au contraire, on connaît deux espèces d'enquêtes, les unes ne contiennent que des chiffres, additionnent le nombre de réponses affirmatives et négatives, présentent en tableaux leurs résultats, — d'autres sont d'expression purement littéraire et ce ne sont pas les moins instructives. Pour ceux qui considèrent la notation numérique des observations comme un caractère secondaire de la statistique, la différenciation ci-dessus ne paraîtra pas bien apparente, mais dans l'opinion que nous avons exprimée plus haut, cette question est au contraire essentielle.

Une seconde différence se marque quand on aborde le « comment » de l'enquête. Si les questions envisagées par la statistique et par les enquêtes sont les mêmes, elles ne sont pas traitées de même. Les bulletins des statistiques sont sobres, concis, peu chargés de détails; ils peuvent suffire à donner une vue d'ensemble du sujet, à en marquer les limites, à mesurer les grandeurs, à faire apparaître les relations causales, mais on a renoncé d'avance, en les dressant, à faire l'inventaire de chacun des traits particuliers des unités. Quelle souplesse et quelle variété, au contraire, dans les questionnaires des enquêtes! On sent que, plus à l'aise parce qu'il opère sur des groupes moins nombreux et plus faciles à manier, l'auteur de l'enquête a pu multiplier les points de vue, entrer dans des détails interdits à celui qui s'adresse à des centaines de milliers de personnes. L'enquête dispose donc d'un questionnaire plus détaillé et plus riche que la simple statistique. Toutefois cette différenciation n'est pas absolument essentielle, car il y a des enquêtes fort concises et des statistiques très détaillées; le bulletin du recensement belge des industries et métiers de 1896 for-

mait, par exemple, une petite brochure d'une vingtaine de pages !

Nous touchons au nœud de la question lorsque nous examinons la question des limites imposées à l'un et à l'autre genre de recherches. Pour exprimer les limites du relevé direct, on a eu recours à une formule extrêmement large : la statistique doit étendre ses investigations au territoire entier, sans en laisser une parcelle inexplorée, et elle doit comprendre dans ses calculs toutes les unités existantes, sans en laisser échapper une seule. Or, aucune exigence de ce genre n'est formulée pour l'enquête ; les limites d'espace sont, en ce qui la concerne, larges et flottantes ; elles s'étendent et se resserrent au gré des circonstances. Dans ces limites mêmes, il n'est pas question de saisir toutes les unités, mais seulement d'en relever un certain nombre prises comme types. La statistique détermine les dimensions générales des phénomènes, l'enquête en précise les contours.

En résumé, l'enquête est un procédé différant de la méthode statistique, mais par son but, son objet et son aspect général, elle se rapproche tellement du relevé statistique proprement dit qu'elle en forme en quelque sorte une partie intégrante. D'après Salvioni, enquête et statistique sont des divisions coordonnées de l'investigation des phénomènes sociaux, seulement l'enquête est essentiellement descriptive et ne vise qu'à donner des résultats approximatifs, alors que la statistique s'efforce de faire apparaître les dimensions des faits observés et fournit les résultats de ses recherches sous la forme numérique.

97. L'enquête, d'une bien plus grande souplesse que la statistique, peut viser une infinité de sujets. Pour s'en faire une idée, le lecteur pourra consulter la liste des publications des différents Offices du Travail en Europe et des Bureaux de statistique du Travail aux Etats-Unis ; des exemples en seront analysés avec détail dans la partie de cet ouvrage consacrée à la statistique du travail (tome III).

Nous distinguons, parmi les enquêtes, plusieurs types appartenant à des classes méthodologiques différentes :

1° L'enquête orale et l'enquête écrite. Les enquêtes parlementaires ou publiques sont généralement des enquêtes orales, consistant en une suite de questions et de réponses suivant les indications d'un programme général. Les documents écrits sont généralement rejetés en annexe et ne présentent qu'une importance secondaire.

Les enquêtes se rapprochant du type de la statistique sont des enquêtes écrites;

2° L'enquête par questionnaires répandus dans le public à un grand nombre d'exemplaires, chacun étant libre d'y répondre ou non; l'enquête par questionnaire à des têtes choisies, individus ou institutions; celle dont les documents sont présentés et repris par des agents du gouvernement; celle enfin où la documentation est recueillie par des agents spéciaux, visitant les endroits désignés et rédigeant les observations sous leur propre responsabilité?

3° L'enquête dont le questionnaire réclame surtout des renseignements ayant un caractère numérique — celle dont le questionnaire comporte des réponses d'expression littéraire — celle sans questionnaire, n'utilisant qu'un programme général d'investigations que l'enquêteur suit du plus près possible.

98. Nous reproduisons, ci-après, une partie d'un questionnaire d'enquête à caractère statistique.

RÉPUBLIQUE FRANÇAISE (1).

Ministère du Travail
et de la Prévoyance sociale.
—
Statistique générale
de la France.
—
Paris, 97, quai d'Orsay.
—
Salaires, durée du travail de certaines catégories d'ouvriers.

Paris, le 3 janvier 1911.

MONSIEUR LE PRÉSIDENT,

J'ai l'honneur de vous demander de bien vouloir remplir et me retourner ensuite, le questionnaire ci-contre, relatif aux salaires ordinaires de quelques catégories d'ouvriers.

Aux catégories désignées, et qui ont été inscrites pour permettre des comparaisons avec les enquêtes antérieures, vous pourrez utilement ajouter celles qui, dans la localité, comportent généralement une certaine tarification des salaires.

Je vous serai, en outre, obligé de bien vouloir indiquer, pour les catégories les plus nombreuses, le prix habituel de pension payé par l'ouvrier seul, afin de permettre de comparer le coût de la vie aux différents points du territoire avec les chiffres qu'ont fait connaître les enquêtes précédentes.

Veuillez agréer, Monsieur le Président, l'assurance de ma considération la plus distinguée.

Le Ministre du Travail
et de la Prévoyance sociale,
L. LAFERRE.

A Monsieur le Président du Conseil de prud'hommes.

---

(1) Extrait de *Salaires et coût de l'existence à diverses époques jusqu'en 1910*, publication de la « Statistique générale de la France ». Paris, 1911, pp. 15-17.

CATÉGORIES d'ouvriers	Nombre annuel de journées de travail pour la majeure partie des ouvriers de chaque catégorie.	Durée la plus habituelle de la journée de travail		Salaire ordinaire de l'ouvrier non nourri		OBSERVATIONS Indiquer, quand il y a lieu, le salaire de l'ouvrier nourri ; indiquer s'il est en même temps logé. Dire si le salaire de l'ouvrier nourri ou logé est payé à la journée ou au mois.
		En été	En hiver	Par heure de travail	Par journée de travail	
Ouvriers :						
Ouvrier agricole . . .						
Garçon jardinier . . .						
Garçon boulanger (brigadier)						
Meunier						
Garçon boucher (étalier au détail)						
Garçon charcutier . .						
Ouvrier brasseur. . .						
Imprimeur-compositeur						
Relieur						
Ouvrier tanneur . . .						
Sellier bourrelier . .						
Cordonnier						
. . . .						
. . . .						
. . . .						
Ouvrières :						
Repasseuse						
Couturière en robes. .						
Lingère						
. . . .						
. . . .						
. . . .						
Autres spécialités importantes dans la région :						
. . . .						
. . . .						
. . . .						

Quelles sont les catégories d'ouvriers du sexe masculin les plus nombreuses dans la région, parmi celles désignées ci-dessus ?

Quel est le prix de pension que payent le plus ordinairement les ouvriers célibataires de ces catégories pour le logement et la nourriture ?

99. Une organisation des enquêtes beaucoup moins étroite que la précédente, consiste à charger du travail de recherche, des agents spécialement désignés, appartenant au personnel administratif ou choisis en dehors, en leur prescrivant uniquement de suivre les indications d'un programme général dressé à leur intention. L'enquête belge sur les industries à domicile, commencée en 1898, qui a paru en dix volumes, appartient à ce type. Nous reproduisons ci-après le début du programme général; on le trouvera *in extenso* dans le premier volume de cette publication; à raison de sa longueur, il est impossible de le donner ici en entier :

Généralités.

Le milieu physique.	Influence des conditions physiques sur la localisation de l'industrie. Topographie générale, le sol, l'air, les eaux.

Le milieu démographique. — Population; sa densité.

Le milieu économique.	Caractère industriel de la région : grande, moyenne, petite industrie. Caractère agricole de la région : grande, moyenne, petite culture Commerce de luxe, commerce d'objets de première nécessité.
Le milieu social et moral.	Etat de perturbation ou de tranquillité sociale. Influence de la coutume et de la tradition. Socialisme, ses moyens d'action et de propagande, leur succès auprès des ouvriers de l'industrie à domicile. Les notions morales, leur influence sur la population travaillant à domicile.

I. — *Organisation commerciale.*

Origine de l'industrie à domicile.	Transformation de la condition des petits patrons indépendants. Régression de l'organisation du travail en fabrique dans le but, par exemple, de réduire les frais généraux.	
Evolution de l'industrie à domicile.	Date et circonstances de l'établissement de l'industrie dans la région. Phases successives de son développement et de ses transformations. Evolution de la technique et de l'outillage. Description sommaire des procédés et du mode de travail actuel.	
	Examen critique des procédés et du mode actuels de travail	au point de vue de la quantité produite; id. de la bonne exécution du travail; id. de l'apprentissage et du recrutement du personnel; id. de la dispersion et de l'agglomération du personnel ouvrier, etc., etc.

Un programme de ce genre est une sorte de table des matières soumise aux enquêteurs et dont ceux-ci ont, le cas échéant, à remplir certaines divisions tout en en passant d'autres sous silence.

100. On a cherché maintes fois à établir des relations systématiques entre le relevé partiel, réalisé au moyen de l'enquête et le relevé complet qu'on pourrait obtenir par un recensement. Le problème consiste à observer des parties proportionnelles au tout, comme nombre, qualité, localisation, etc. Question ardue et complexe, supposant toujours, à la base, un relevé général préalable que l'on prend pour guide et comme moyen de vérification. L'exposé systématique des enquêtes représentatives n'existe pas; cependant, toutes les institutions dont l'enquête est le principal moyen d'investigation, ont toujours essayé de rendre leurs recherches le plus représentatives possible. Le nombre des unités comprises dans l'enquête n'est pas, loin de là, le seul élément à considérer; il faut encore que toutes les classes existantes dans la masse se trouvent représentées proportionnellement et que les observations soient réparties sur toute l'étendue du territoire. Des essais intéressants ont été réalisés en Norvège, en 1894; M. Kiœr, directeur du bureau central de statistique de ce pays, en a donné un exposé complet dont nous résumons ci-après quelques points principaux. La question qui a donné lieu à l'enquête représentative exécutée en Norvège, était celle de la création d'une caisse générale de retraite et d'assurance contre l'invalidité et la vieillesse. Il ne pouvait être question de recourir à un dénombrement général à raison du peu de temps dont on disposait et aussi du nombre considérable des questions posées par le questionnaire (plus de soixante).

Un premier élément de différenciation est la distinction entre les villes et la campagne. Le nombre des individus adultes à interroger par le moyen du dénombrement représentatif étant fixé à 80,000, on répartit ces bulletins, d'après

la proportion établie par le précédent recensement, entre les villes et les campagnes, soit 20,000 pour les villes et 60,000 pour les campagnes. Un second élément se remarque aussitôt : les grandes villes, les villes moyennes et les petites villes. Il a fallu désigner les localités urbaines dans lesquelles s'effectuerait le dénombrement : on en a pris, réparties entre ces trois catégories, treize en tout, représentant environ le cinquième des villes du royaume.

L'attribution des bulletins réservés aux villes (20,000), d'après le chiffre de leur population, aurait eu pour conséquence de faire remplir un trop grand nombre de bulletins dans les grandes villes : on a donc donné aux villes moyennes et petites un nombre de bulletins proportionnellement plus élevé.

Pour la répartition des bulletins à l'intérieur de la plus grande ville du royaume, Kristiania, on a suivi un système assez curieux : les rues de la capitale ont été classées d'après leur population, puis on a procédé comme suit : on a recensé la population adulte entière de 1/20 des rues les moins populeuses, puis celle de 1/10 des rues de la seconde catégorie, une maison sur deux; ensuite celle de 1/4 des rues de la troisième catégorie en prenant une maison sur cinq; enfin, celle de la 1/2 des rues les plus populeuses mais où seulement une maison sur dix devait être visitée par les agents de l'enquête.

L'organisation du dénombrement représentatif s'est inspirée de règles analogues dans les campagnes, comme on pourra le voir en recourant au mémoire original de M. Kiœr. D'après les chiffres produits par le statisticien norvégien, le rapport des diverses classes d'unités comprises dans l'enquête et celui des mêmes catégories trouvées à l'aide du recensement général concordent d'une manière fort satisfaisante (1).

(1) A. N. Kiœr, « Observations et expériences concernant des dénombrements représentatifs ». *Bulletin de l'Institut international de statistique*, t. IX, 2ᵉ livraison, p. 176.

II. — La monographie.

101. Si l'on s'écarte de la notion de la statistique avec l'enquête, on s'en éloigne bien davantage avec la monographie. L'idée de phénomène collectif est encore présente dans les recherches organisées sous la forme de l'enquête ; celle-ci est une statistique moins complète que le relevé direct, mais les faits qu'elle considère sont de la même essence. La monographie, au contraire, s'attache à un objet qu'elle détaille *con amore,* elle en analyse toutes les parties, elle en scrute les derniers replis. Mais, par le fait même qu'elle abandonne la recherche collective pour se vouer à l'observation individuelle, la méthode monographique s'éloigne du domaine de la statistique. On n'a pas toujours appliqué à ces notions différentes une critique assez sévère, ce qui fait qu'on a confondu des genres que la logique doit séparer. Le fait seul que la monographie s'acharne à la recherche du « type » n'est-il pas suffisant à la faire classer en dehors de la statistique, laquelle ne concerne que les phénomènes non typiques ? Cheysson a déployé beaucoup de diplomatie pour faire admettre la monographie au rang d'une sœur cadette de la statistique. Nous ne pouvons cependant être d'accord avec lui que le relevé direct ou statistique proprement dit jouerait dans l'ensemble de la méthode le rôle de la synthèse, et que celui de l'analyse serait réservé à la monographie. En réalité, la statistique se limite au relevé complet des unités et, par extension, au relevé partiel, par voie d'enquête ; elle ne peut consister en la description minutieuse d'un seul fait, fût-il qualifié de fait-type.

Nous résumons ainsi les différences essentielles existant entre la méthode statistique proprement dite et la méthode monographique :

1° La statistique est une méthode propre à l'observation des phénomènes collectifs, qui étudie les faits variables afin d'arriver à la notion du caractère normal. Il

en résulte qu'elle considère les masses, les faits collectifs dans leur ensemble. La monographie ne se propose pas le même but et loin de se livrer à l'étude des masses elle se borne avec soin à l'examen d'un fait soigneusement délimité ;

2° La statistique et la monographie ne parlent pas le même langage. Celui de la statistique est essentiellement numérique. Celui de la monographie est littéraire et descriptif ; il comprend des aperçus historiques, philosophiques, économiques, sociologiques et les chiffres n'y interviennent qu'à titre de documentation ;

3° Les méthodes mathématiques de la statistique, employées dans le but de dégager les résultats généraux, telles que les moyennes, les mesures de la dispersion, l'étude des corrélations, etc., ne trouvent pas d'application dans les recherches monographiques pour la raison qu'elles supposent l'existence d'un très grand nombre de faits, alors que la monographie n'en étudie qu'un seul ou qu'un petit nombre.

La monographie a été utilisée par Le Play pour la description des familles et complétée par l'établissement du budget de dépenses et de recettes. On a appliqué le même procédé descriptif à l'atelier et même à la commune. On peut trouver dans ces travaux des traits de ressemblance entre les deux méthodes statistique et monographique ; ainsi, le domaine qu'elles explorent l'une et l'autre est presque identique ; ce sont les questions se rapportant aux phénomènes sociaux et économiques. On peut traiter les questions de la natalité, de la nuptialité, des salaires, des accidents du travail par la méthode monographique comme par la méthode statistique. Mais l'identité du domaine exploré n'implique pas l'identité du procédé de recherche. Une autre raison invoquée est que les deux méthodes s'efforcent de comparer leurs résultats dans l'espace et dans le temps ; il nous semble que ce but est commun à toutes les recherches scientifiques et ne suffit pas à conclure à l'iden-

tité (1). Quelle que soit la valeur des œuvres conçues sur ce type, on ne peut, pour des raisons de logique et de méthode, les comprendre dans le domaine de la statistique.

102. *Références :*

BENINI (R.), *Principii di statistica metodologica.* Torino, 1906, pp. 37-59.

Id., *Statistica metodologica e statistica economica, Lezioni dettate all' Università di Roma,* 1910-11. Roma, 1911, pp. 27-36.

BERTILLON (J.), *Cours élémentaire de statistique administrative.* Paris, 1895, pp. 43-68.

BLODGETT (James H.), *Obstacles to accurate statistics* (American statistical Association, 1898-99), p. 1.

BOSCO (A.), *Lezioni di statistica. Parte prima. Metodologia statistica.* Roma, 1909, pp. 151-215.

BOWLEY (A. L.), *Elements of statistics.* Second edition. London, 1902, pp. 17-20.

Id., *An elementary manual of statistics.* London, 1910, *passim.*

COLAJANNI (N.), *Lezioni di statistica.* Napoli, 1903, pp. 34-58.

DURAND (E.), *Changes in census methods for the census of 1910* (American statistical Association, 1910-11). Boston, 1912, pp. 53-66.

FAURE (F.), *Eléments de statistique.* Paris, 1906, pp. 67-71.

GABAGLIO (A.), *Teoria generale della statistica.* Milano, 1888. t. II, pp. 63-94.

GAIT (E. A.), *Census of India,* 1911, Part. I Report. Calcutta, 1913. Introduction, pp. V-XI.

JULIN (A.), *Précis du cours de statistique générale et appliquée,* quatrième édition. Bruxelles, 1919, pp. 17-38.

KING (W.), *The elements of statistical method.* New-York, 1912, pp. 39-60.

LEVASSEUR (E.), *La population française,* t. I. Introduction sur la statistique. Paris, 1889, p. 27.

LIESSE (A), *La statistique, ses difficultés, ses procédés, ses résultats.* Paris, 1905, pp. 21-45.

MARCH (L.), *Statistique* (de la méthode dans les sciences, deuxième série). Paris, 1911.

MAYR et SALVIONI, *La statistica e la vita sociale.* Seconda edizione. Torino, 1886.

---

(1) CHEYSSON, « La monographie d'atelier ». *Bulletin de l'Institut international de statistique,* 2e session, 1er fascicule, p. 92.

Mayr (G. von), *Statistik und Gesellschaftslehre.* Erster band : Theoretische statistik, Zweite auflage. Tübingen, 1914, pp. 65-99.

Meitzen (A.), *History, theory and technique of statistics.* Trad. anglaise de Roland P. Falkner. Philadelphia, 1891, pp. 110-123.

Reports of the Department of Commerce and Labor (U. S. A.), Bureau of the Census. Washington, 1912, pp. 45 et suivantes.

Quetelet (A.), *Lettres sur la théorie des probabilités.* Bruxelles, 1846, pp. 288-297.

Tammeo (G.), *La statistica.* Torino, 1896, pp. 101-109.

Verryn-Stuart, *Inleiding tot de beoefening der statistiek.* Harlem, 1910, pp. 30-34.

Virgilii (F), *Statistica,* cinquième édition. Milano, 1911, pp. 37-49.

Weston (Stephens, F.), *Limitations of statistics* (American statistical Association, 1892-93), p. 259.

SECTION III

La critique statistique

CHAPITRE PREMIER

Généralités, définition, division

I. — Degré de précision des résultats statistiques.

103. — Supposons achevées les opérations du relevé : les agents recenseurs sont allés de maison en maison, y ont laissé autant de bulletins qu'il le fallait, se sont présentés à nouveau pour les recueillir, ont vérifié les réponses inscrites au bulletin, les ont complétées ou les ont eux-mêmes rédigées, si les recensés en étaient empêchés. Le dénombrement a été exécuté dans toutes les communes, ses opérations répondent à la condition de synchronisme indispensable en cette matière, les questions du bulletin ont été en général bien comprises, les administrations subordonnées ont rempli avec conscience leurs obligations, bref le relevé s'est effectué — admettons-le — dans des conditions aussi satisfaisantes qu'il est possible de le souhaiter... Est-ce à dire que l'on puisse utiliser immédiatement, sans retouches, le matériel tel qu'il vient d'être réuni ?

Avant de répondre à cette question, il convient d'examiner deux problèmes de nature générale : *a*) celui des limites de la précision en matière de relevé statistique ; *b*) celui du degré de centralisation des opérations.

A. — Dans les sciences, la précision des mesures calculées atteint aujourd'hui un degré extrêmement élevé. Les lunettes et les télescopes astronomiques, par exemple, sont, comme le dit M. Edmond Bouty, les plus parfaits des instruments de précision. « Ils dépassent de bien loin tous les instruments de mesure employés dans les laboratoires de physique. Ce degré d'exactitude est imposé par l'état présent de l'astronomie. Dans bien des cas, une erreur d'une seconde d'arc serait considérée comme intolérable. Les mesures courantes sont faites au dixième de seconde d'arc, et l'on se rendra compte de ce que peut être une telle précision, si l'on rappelle que l'arc d'un dixième de seconde mesuré à la surface de la terre n'est que de trois mètres environ (1) ».

Cette même précision, nous la rencontrons jusque dans les sciences appliquées ; il n'est pas rare de voir prendre des mesures, à l'aide d'un vernier, au dixième de millimètre, et il n'est pas extraordinaire de voir cette précision poussée jusqu'au vingtième de millimètre. De simples ouvriers sont parfois chargés de vérifications portant sur des dimensions aussi imperceptibles.

104. Il n'est pas question dans les recherches statistiques, d'arriver à une précision comparable à celle-là. Nous en sommes loin de compte et pour cause. Que l'on songe d'abord à ce que le relevé comporte, par lui-même, d'erreurs, d'omissions et de multiples emplois ; ensuite, à la difficulté des définitions, à la grossièreté des moyens de compte, à l'indolence de ceux qui livrent le renseignement et de ceux qui le recueillent, et l'on sera édifié. Le relevé statistique donne, par rapport à la vérité absolue, une simple approximation. Pour en être convaincu, il suffit presque de poser les termes du problème ; la démonstration, souvent, paraîtra superflue.

(1) *La vérité scientifique, sa poursuite,* par Edm. Bouty. Paris, Ern. Flammarion, 1908, p. 172.

Voici un grand port de mer, dont la statistique relève le mouvement, en même temps qu'elle indique les quantités de marchandises débarquées et embarquées : il s'agit d'Anvers. Nous voyons, entre autres, qu'en 1912, on y a débarqué les quantités suivantes de marchandises : peaux brutes, 57,970,501 kilogrammes ; os et cornillons, 35,168,958 kilogrammes ; guano, 12,861,589 kilogrammes. Cette précision, pour impressionnante qu'elle soit, ne laisse pas de nous causer quelque inquiétude. Quelle vraisemblance y a-t-il que ces marchandises pondéreuses, encombrantes et malpropres aient été minutieusement pesées, à mille grammes près ? Dans ces conditions, nous aimerions mieux qu'on s'en tînt à la tonne, c'est-à-dire que la précision fût mille fois moindre ; on s'estimerait encore trop heureux d'avoir des mesures aussi précises et peut-être même notre sécurité serait-elle plus complète si on se bornait à une approximation encore plus grossière. Les statistiques minières qui font connaître le chiffre de la production, y mettent plus de discrétion ; elles se bornent à indiquer le nombre de tonnes extraites. Personne ne croira impossible cependant qu'un certain nombre de berlaines soient comptées en trop ou en moins et qu'une erreur puisse provenir de là, comme elle peut avoir pour cause une approximation par trop fruste des quantités déjà extraites, existant en stock. Supposons qu'on veuille aller plus loin, qu'on ait l'ambition d'appliquer au relevé des mesures plus précises. Il faudrait commencer à mesurer exactement le contenu de chaque berlaine, chose très compliquée si nous entendons appliquer les procédés scientifiques qui sont de mise en matière de mesure. Il faudra reconnaître bientôt que de tels procédés sont impossibles à appliquer et surtout sont tout à fait inutiles, parce que la quantité totale de combustible extrait est si considérable qu'une erreur même importante en elle-même, est tout à fait négligeable par rapport à la masse. Le bon sens doit rester notre maître en toute chose, et particulièrement en statistique. Une erreur d'un gramme, d'un cen-

tigramme, d'un milligramme et moins, pourra être tout à fait intolérable en chimie et avoir des conséquences désastreuses; une erreur de mille, de dix mille ou de cent mille kilos est fort indifférente en bien des matières dont la statistique s'occupe. Le résultat du relevé statistique n'est et ne peut être qu'une approximation.

Cette analyse peut paraître un peu superflue; il n'en est rien, elle doit être prolongée encore pour dégager définitivement des notions claires et justes. Dans un recensement de la population, on compte le nombre de personnes se trouvant, à une date donnée, à un endroit déterminé et le nombre de personnes qui y ont légalement leur résidence. Il est permis de se demander si ce compte, à une personne près, a des chances d'être exact? Il y a deux espèces d'exactitude; la première : la conformité du calcul avec le document, ce qu'on veut bien admettre, par une hypothèse toute gratuite, d'ailleurs... il est certain que dans le comptage il s'est produit des erreurs. La seconde espèce d'exactitude c'est le rapport du relevé à la réalité. Comment supposer qu'aucune omission n'a été commise; qu'aucun bulletin n'a été égaré; qu'aucune personne, dans les grandes casernes qui servent d'habitation à des dizaines de ménages, n'a omis de remplir son bulletin et que l'agent recenseur s'est aperçu de cette lacune? On peut en dire tout autant du recensement industriel qui relève le nombre d'entreprises, de patrons, d'ouvriers, de moteurs, etc. Et là, un danger plus grave menace l'exactitude du relevé : c'est l'omission d'une unité importante et la répétition de la même unité déclarée par plusieurs personnes ayant un titre quelconque à faire cette déclaration. En Belgique, lors du recensement industriel de 1896, une fabrique comprenant plusieurs milliers d'ouvriers, fut recensée trois ou quatre fois, en même temps, par suite des déclarations émanées du directeur, de l'ingénieur, de l'administrateur délégué, etc. Hâtons-nous d'ajouter que la critique instituée eut pour résultat d'empêcher ce comptage multiple.

Qu'il s'agisse d'un comptage de choses ou d'un relevé de personnes, la statistique ne peut donner de certitude absolue si l'on recherche un degré de précision tant soit peu élevé. C'est du reste d'une parfaite inutilité dans un grand nombre de cas.

105. Cette conclusion ne peut manquer de provoquer une objection, sinon dans l'esprit du lecteur, du moins parmi les sectateurs d'une sagesse vulgaire et un peu épaisse : « Si la statistique, diront-ils, ne peut donner que des approximations, est-il bien nécessaire de s'acharner à la rendre plus exacte ou plus précise qu'elle n'est quand elle arrive en nos mains? A quoi bon substituer une erreur un peu moins grande à une erreur un peu plus forte? Est-on bien certain que la seconde mesure sera plus exacte que la première? » Admettre un tel postulat équivaudrait, en réalité, à condamner tout le mouvement scientifique contemporain. De ce que les anciens instruments astronomiques ne donnaient qu'une approximation assez grossière, à cause de leur faiblesse et de la réfraction provoquée par leurs lentilles, fallait-il conclure à l'inutilité de recherches nouvelles? Les calculs basés sur ces mesures insuffisantes étaient-ils absolument inutiles? Penser que la statistique peut, sans inconvénient, se contenter de prendre tel quel le matériel qui lui est apporté par l'observation primaire, c'est la condamner, de dégradation en dégradation, à un empirisme vide de toute pensée scientifique. L'énoncé de pareille conséquence dispense d'une réfutation plus étendue.

II. — Influence de la centralisation sur l'exactitude des résultats.

106. Le lecteur se rappelle notre division du relevé statistique : relevé direct ou complet, relevé indirect ou partiel. Pour plus de clarté, envisageons seulement le relevé direct, celui auquel on peut, à juste titre, donner le nom de relevé statistique. Les bulletins se trouvant remplis, il faut les dépouiller et en faire la critique. Mais entre les deux opéra-

tion : le relevé d'un côté, la critique et le dépouillement de l'autre il existe un hiatus que le lecteur aura déjà aperçu : comment s'opèrent la réunion, le classement et la transmission à l'organe central des centaines de milliers de bulletins recueillis sur tous les points du territoire ?

Divers systèmes sont en présence ; le choix à faire n'est pas indifférent sous le rapport de la possibilité de l'application des règles critiques. Le plus ancien pourrait être appelé : système de la décentralisation. Dans l'organisation du dépouillement on a pris soin, dans ce cas, de décharger autant que possible l'organe central de toute besogne d'exécution pour ne lui laisser à accomplir qu'un travail de coordination. La description de cette organisation, ainsi que de celle dite de la « centralisation » trouvera sa place à l'endroit où il est traité du dépouillement. Nous n'avons, pour le moment, qu'à la considérer sous le rapport des facilités ou des obstacles qu'elle présente pour la critique statistique.

A cet égard, la valeur d'une statistique exécutée dans ces conditions nous semble très faible, pour ne pas dire davantage.

La critique ne se trouve organisée nulle part. Si elle peut se faire lors d'une phase quelconque du travail, ce ne peut être que lors du second passage de l'agent recenseur, lorsqu'il vient recueillir les réponses des recensés. Il lui est prescrit de s'assurer alors de l'exactitude des mentions portées au bulletin, mais dans quel sens entend-on ce terme « exactitude » ? Prenons comme exemple un recensement de ce type, d'ailleurs parfaitement organisé, si l'on admet la justesse du point de départ, et voyons ce que les instructions aux agents recenseurs contiennent de règles positives relatives à la critique du document. La vérification dont il s'agit prend un caractère administratif, dès la définition qui en est donnée : « le soin de veiller à ce que tous les renseignements consignés dans chaque bulletin le soient *conformément aux prescriptions du gouvernement* est l'acte le plus important dont l'agent recenseur ait à s'acquitter dans l'exercice de ses fonctions. »

Ce contrôle revêt ainsi un caractère assez formaliste. L'agent recenseur est invité, il est vrai, à s'assurer si toutes les personnes qui doivent figurer au bulletin s'y trouvent mentionnées, si l'on n'y a point inscrit des personnes qui ne devraient pas être comprises dans un document de ce genre, etc. Les instructions passent ensuite en revue une série de cas : que fera l'agent recenseur dans telle circonstance? Comment doit-il agir dans telle autre? Il est dit que la mention de tout pseudonyme ou sobriquet est interdite, que tous les prénoms doivent être énoncés si c'est possible dans l'ordre qui leur est assigné dans l'acte de naissance; nous voyons aussi que les jeunes enfants, qui ne sont pas encore en âge de parler, sont inscrits sur le bulletin comme ne parlant « aucune langue », tandis que les muets sont censés parler la langue ou les langues nationales dont ils se servent habituellement pour exprimer leurs idées. Mais toutes ces règles, pour bien fondées qu'elles soient, ne constituent pas la critique du document. Nous pensons y arriver en entamant le chapitre important des « professions, fonctions ou positions ». Notre attente est déçue, car à propos de la question si difficile des distinctions économiques entre patrons, employés et ouvriers, nous ne trouvons que cette brève mention : « le recensé qui exerce une profession industrielle, commerciale ou agricole, mentionnera s'il l'exerce comme maître, patron ou chef d'exploitation, comme employé ou surveillant, comme ouvrier ou manœuvre. »

« Les qualifications souvent usitées de journalier, ouvrier, manœuvre, ne doivent être employées qu'en spécifiant, en même temps, le genre de métier exercé par l'intéressé. Exemples : journalier agricole, ouvrier terrassier, manœuvre de maçon, manœuvre de houillère, etc. (1) »

En dehors de ce moment, d'après la reprise des bulletins,

(1) Belgique : Recensement général de la population au 31 décembre 1910. *Carnet d'instructions*, ch. III, 1.

l'agent recenseur n'aura plus l'occasion de contrôler quoi que ce soit : absorbé par des besognes administratives, il emploie son temps à remplir des formules, à aligner des chiffres, mais il se trouve dans l'impossibilité de vérifier les déclarations faites par les recensés. L'organe central en est plus empêché encore, puisqu'il n'est en possession d'aucun document original, mais seulement de tableaux récapitulatifs numériques. Sans doute, si l'une ou l'autre erreur énorme est commise dans ces dépouillements préliminaires, l'organe central pourra s'en apercevoir. Mais il ne pourra jamais rectifier les fausses déclarations en vertu desquelles, par exemple, des patrons se sont inscrits comme ouvriers, ou vice-versa. Sans le document original mûrement examiné, pesé, tourné et retourné, il n'y a pas de critique possible. La centralisation des opérations consécutives du relevé est donc la condition d'une critique efficace. A supposer même qu'on puisse en confier la critique à des agents recenseurs d'une qualité très supérieure à ce qui est habituel, le nombre même de ces agents rendrait difficile un pareil travail : les règles tracées n'auraient guère chance d'être appliquées d'une façon uniforme. Il faut donc, non seulement, travailler sur le document original, mais, de plus, c'est au centre que la vérification doit s'opérer.

III. — Nécessité de la critique statistique.

107. Le développement de la critique a toujours marché de pair avec celui de l'esprit scientifique. Le jour où le premier savant a soumis à une enquête méthodique les éléments de ses connaissances, la critique est née et, avec elle, l'esprit scientifique. La statistique n'a pas fait exception à la règle. Ce serait une recherche de pure érudition d'essayer de trouver quel est celui, parmi les nombreux professeurs de statistique et auteurs de manuels, qui eut le premier l'idée de la valeur de la critique. Remarquons-le en passant : cette idée est très nettement exprimée dans les écrits de celui qui,

précisément, a imprimé à la statistique un caratère scientifique ; Quetelet, dans ses « Lettres sur la théorie des probabilités » y a consacré un paragraphe (1). La critique est d'autant plus nécessaire dans notre domaine que les chiffres impressionnent davantage le lecteur confiant. Dans un volume de format imposant, de sévères colonnes de chiffres donnent l'impression de recherches minutieuses, de résultats précis, de conclusions nettes, certaines. Le chiffre est fascinateur, comme l'a très bien montré Seignobos (2). Et il faut un véritable effort, parfois, pour s'arracher à la séduction qu'il exerce, à l'impression de sécurité qui émane de lui. En histoire, en philosophie, nous savons bien que nous sommes exposés à subir l'influence de l'auteur ; mais dans l'exposé qu'il nous fait, si nous avons quelque peu l'esprit critique, nous démêlons bien vite ses tendances, sa façon de travailler, ses préoccupations secrètes, ses convictions religieuses et philosophiques, et contre tout cela nous nous tenons en garde.

108. En statistique, rien de pareil. La personnalité de l'auteur n'apparaît pas. Elle est muette et sans accent et cependant cette personnalité, non seulement existe, mais elle existe à des milliers d'exemplaires. Il n'y a pas que celle de l'éditeur de l'ouvrage, il y a les personnalités multiformes de tous ceux qui ont répondu au bulletin. Quand l'historien analyse un document, il l'examine sous tous les points de vue, il le lit et le relit, il scrute les mobiles, il analyse les passions, il en recherche les manifestations dans l'expression littéraire adoptée par l'auteur du document. Il étudie ainsi un, dix, cent documents. Le statisticien en a des centaines de milliers devant lui qui, eux aussi, méritent une analyse. N'exagérons rien ; ce sont

(1) Quetelet, *Lettres sur la théorie des probabilités.* Bruxelles, Hayez, 1846, p. 298.

(2) Seignobos, *La méthode historique appliquée aux sciences sociales.*

des documents d'une nature plus simple, moins étendus que les documents historiques proprement dits. Et puis, la plus simple technique a enseigné le moyen de surprendre, en flagrant délit, le menteur intentionnel. La besogne est moins ardue, mais elle est nécessaire en statistique comme en histoire, car un bulletin statistique est un document et qui dit document dit un écrit, un témoignage, où viennent se refléter les passions qui agitent les hommes.

La fonction de la critique statistique consiste donc tout d'abord à éliminer des réponses les éléments douteux introduits par l'intérêt, la défiance, ou simplement la paresse. Puis, se souvenant qu'une organisation compliquée ne peut fonctionner sans quelques accrocs, le statisticien s'efforcera de corriger les erreurs matérielles provenant des omissions et des doubles emplois dans le relevé, des fautes commises dans le comptage, et d'effacer jusqu'aux erreurs matérielles de calcul, d'écriture et d'impression.

De là, une division toute naturelle de la matière : la critique de sincérité qui s'applique aux éléments intentionnels; la critique d'exactitude qui vise les cas d'erreur ayant une origine accidentelle.

109. Les erreurs que constate la critique sont : 1° constantes, ou 2° accidentelles. Les premières sont celles qui se reproduisent autant de fois que l'observation est répétée et demeurent les mêmes lorsque les circonstances de l'observation restent identiques. Les défauts d'organisation du relevé, les dispositions psychologiques de l'observateur ou du sujet, des instructions peu claires créent des erreurs constantes.

Les erreurs accidentelles (ou irrégulières) sont celles produites par une quantité de petites causes qui troublent l'observation tantôt sur un point, tantôt sur un autre et qui, à cause de leur complexité, ne peuvent être connues à l'avance. Une bonne organisation statistique ne peut les empêcher toutes, mais en réduit le nombre; elles se distri-

buent selon une certaine loi et sont d'autant moins opérantes que le nombre des observations est élevé.

Les erreurs constantes sont parfois « intentionnelles » à raison des dispositions psychologiques de l'auteur du document : telle est la tendance à déclarer, lors des recensements, un âge arrondi à la décimale la plus proche.

Cette distinction correspond à celle que l'école anglaise désigne du nom de « Biassed » et « Unbiassed errors », distinction à laquelle les écrivains de cette école attachent une grande importance. Pour mieux faire saisir la différence entre les unes et les autres, M. Bowley choisit l'exemple suivant : imaginez que l'on fasse une enquête sur les salaires avec la pensée de montrer que les salaires sont élevés et que l'on emploie des enquêteurs ayant accepté de se conformer à cette tendance. Les enquêteurs se rendront dans différentes localités industrielles et s'informeront des salaires des ouvriers les plus qualifiés et les plus réguliers ; de cette façon les chiffres produits seront trop élevés par rapport à la situation moyenne. Si dans une autre enquête portant sur le même objet, la direction se laisse guider uniquement par une pensée scientifique et veut faire une enquête impartiale, les enquêteurs pourront, il est vrai, choisir des situations qui sont tantôt supérieures, tantôt inférieures à la moyenne, mais le cas est tout différent du premier.

Dans le premier cas, il y aura des erreurs constantes (biassed errors), toutes dans la même direction, tendant toutes à augmenter la moyenne ; dans le second cas, les erreurs seront accidentelles (unbiassed errors) et plus grand sera le nombre des estimations, plus petite sera l'erreur résultant des appréciations fautives.

La différence essentielle entre les erreurs constantes et les erreurs accidentelles consiste, en ce qui concerne ces dernières, en ce que les erreurs se balancent et s'annihilent à peu près, tandis que dans le cas des premières les erreurs se maintiennent au même niveau et s'ajoutent les unes aux autres. Si l'on procède à plusieurs observations (n) qui

donnent toutes la même erreur x, on a comme résultat final la moyenne

$$\frac{x \cdot n}{n} = x$$

grandeur égale à l'erreur constatée dans chaque observation.

110. Les erreurs commises au cours du relevé sont donc : *a*) intentionnelles ; *b*) fortuites. Dans le premier cas, le renseignement donné a été volontairement altéré, soit par celui qui l'a fourni, soit par celui qui l'a recueilli. Les mobiles qui ont déterminé à ne pas répondre conformément à la vérité varient ; on les fixe par l'analyse psychologique. Dans le second cas, l'intention de tromper n'apparaît pas ; des raisons diverses expliquent le mécanisme de l'erreur : insouciance, questions mal comprises, erreur matérielle, etc. Ces causes sont connues par l'analyse des faits et des circonstances. Les erreurs intentionnelles sont synthétisées par cette formule : « On a voulu nous tromper », les erreurs fortuites, par cette autre : « On s'est trompé ». De là, dans la critique statistique, deux divisions bien tranchées : 1° la critique de sincérité, au moyen de laquelle on fait apparaître les causes d'insincérité ; 2° la critique d'exactitude, dont le but est de découvrir les erreurs fortuites. Ces deux parties envisagent respectivement les erreurs constantes et accidentelles dont l'importance relative, très différente, a été définie plus haut.

CHAPITRE II

Critique de sincérité

Mobiles psychologiques influençant la sincérité statistique.

111. Le présent paragraphe se rapporte exclusivement à la critique de sincérité, ou, dans un sens plus étendu, à la recherche des erreurs constantes. Il n'y a pas que le

manque de sincérité qui produise des erreurs constantes : des défauts d'organisation, le manque de clarté des instructions données aux recensés, comme aux agents recenseurs, peuvent conduire à des résultats tout aussi erronés que la crainte ou l'intérêt. Cependant, comme il faut donner un nom à cette partie de la critique, nous l'avons appelée « critique de sincérité » à raison des dispositions d'ordre psychologique qu'elle a en vue, en ordre principal. Il faut s'entendre aussi sur l'expression « erreurs constantes ». On ne veut point signifier par là des erreurs se reproduisant à chaque occasion, mais simplement des erreurs ayant une tendance à se répéter dans un même sens un grand nombre de fois, au lieu que les erreurs fortuites dépassent parfois la mesure et restent parfois en deçà. La crainte et l'intérêt sont des mobiles dont l'action est très étendue, mais il est clair qu'ils ne vicient pas toutes les réponses, tout en en rendant suspectes un grand nombre. Ces distinctions ne doivent pas être perdues de vue au cours de l'exposé qui va suivre.

112. L'inexactitude voulue des réponses données au bulletin statistique provient, très fréquemment, de la crainte des auteurs du document que des réponses sincères ne tournent à leur préjudice.

Cette crainte résulte de l'ignorance ou de l'intérêt mal entendu; il peut s'agir de l'intérêt matériel et de l'intérêt moral. Examinons brièvement les diverses hypothèses qui peuvent se présenter.

A. — *Crainte de mesures fiscales ou de réglementation.* — Mobile très général auquel obéissent de nombreuses personnes. Des patrons, principalement de la petite industrie, hésitent à inscrire sur leur bulletin le nombre exact de leurs ouvriers et de leurs employés, afin de ne pas faire classer leur établissement dans une catégorie plus élevée en vue de la patente. En 1846 déjà, Quetelet signalait cette cause d'inexactitude à propos du premier recensement général de

l'industrie qui eût été, jusqu'alors, exécuté en Europe. L'interprétation fondée sur ce mobile est encore exacte aujourd'hui, bien que le progrès de l'instruction et l'expansion des affaires l'aient atténuée ou même fait disparaître dans une certaine mesure. A notre époque, les grands établissements industriels, groupant dans leurs ateliers et leurs chantiers une fraction de plus en plus notable de la population ouvrière, sont habituellement rangés dans la première classe des patentes, de telle sorte que l'intérêt qu'il y avait à déclarer un nombre d'ouvriers inférieur à la réalité n'existe plus; d'ailleurs, en présence de l'énormité des taxes qu'ils acquittent déjà, la majoration d'un impôt accessoire ne peut plus présenter aux yeux des industriels qu'un intérêt secondaire. Par contre, cet intérêt continue à exister pour les chefs des petites entreprises, plus resserrés dans leurs moyens et plus portés à lésiner. Un critère de différenciation sera donc la grandeur, l'importance de l'entreprise : le statisticien se montrant plus sévère et se tenant plus sur ses gardes à l'égard des déclarations faites pour des exploitations de minime importance que pour celles relatives à des établissements très importants.

On ne pourrait en dire autant de la critique appliquée aux questions du bulletin ayant rapport à des mesures de réglementation du travail. Souvent, l'opposition aux mesures législatives de ce genre est conduite par des associations de grands industriels. Il est à craindre dès lors que les réponses individuelles ne soient inspirées de la même politique et ne procèdent d'un mot d'ordre. Répondre dans un sens indiqué d'avance, et non conformément à la réalité, est un tort qui n'est pas uniquement celui d'industriels non interventionnistes; on peut l'imputer aussi à d'autres milieux où les idées sont au pôle opposé : nous voulons dire les syndicats ouvriers. Que d'enquêtes sans vie, sans portée et sans résultat à cause de ce parti pris! Les réponses arrivent stéréotypées, identiques pour des industries ou des localités placées dans des conditions différentes. Rédigées par un

secrétaire de syndicat, du fond d'un bureau, elles ne s'adaptent même pas aux énoncés précédents et cette unanimité, loin de faire impression, ne fait que provoquer une défiance trop justifiée.

Est-il besoin d'ajouter qu'une enquête viciée dans ces conditions est inutilisable et ne peut même être améliorée? Le parti pris, les réponses formulées d'après un mot d'ordre, fournissent l'exemple le plus complet d'erreurs constantes.

113. Si le statisticien ne peut toujours éviter de poser des questions de nature à faire redouter des répercussions d'ordre financier ou législatif, il peut, tout au moins, prendre des mesures d'abord pour prévenir une crainte légitime, ensuite pour empêcher que les fausses déclarations ne viennent fausser les résultats. On ne manquera pas de faire remarquer aux recensés le but scientifique de la statistique, but totalement étranger aux préoccupations d'ordre fiscal. Le bulletin du recensement belge de 1896 débute par plusieurs « observations générales » dont la première est ainsi conçue : « Le recensement industriel actuel n'est que la répétition d'opérations du même genre effectuées en 1846, en 1860 et en 1880. Il a pour but de faire connaître les principales conditions de l'industrie ; il ne se rattache à aucun projet déterminé de réglementation, ni à aucune mesure fiscale ». Le recensement de 1910 a répété le même avis, en y ajoutant l'exemple du recensement précédent, celui de 1896.

Au dos du bulletin du recensement français de la population figure une « notice sur le dénombrement » précisant la portée, spécialement démographique, de l'opération statistique et faisant ressortir l'ancienneté de ce relevé. Elle est suivie de la note suivante : « Nota — Les réponses que chaque habitant fera, en toute sincérité, aux questions posées ne peuvent, en aucune façon, et à aucun moment, lui occasionner un trouble ou un dommage quelconque ».

S'il est utile de rassurer le public au sujet de la portée

des enquêtes statistiques, il est d'autre part indispensable de prendre des garanties contre le manque de sincérité de ceux des recensés qui ne se seraient pas laissé persuader. Ce sera peut-être le plus grand nombre. Au moyen d'une rapide analyse psychologique, on arrive facilement à déterminer celles des données du questionnaire capables d'éveiller quelque appréhension. Ces données sont ensuite soumises à une épreuve systématique. Il ne suffit pas de rester sceptique devant certaines assertions. Le statisticien doit arriver à une solution positive. Si tel chiffre est supposé inexact, comment le savoir de science certaine, et par quel autre le remplacer? La technique moderne a organisé plusieurs contrôles généraux de ce genre. Au nombre des plus intéressants figure celui qui met en parallèle les déclarations des patrons et celles des ouvriers.

114. On trouve ainsi moyen de corriger des erreurs généralisées en dépit des instructions et des mesures d'organisation. Cette façon de procéder a été adoptée par plusieurs recensements. On a imaginé de faire déclarer par les ouvriers et employés le nom, l'adresse et l'industrie de l'employeur; de cette manière, on dispose d'un matériel de contrôle à l'égard du chiffre indiqué par le patron en ce qui regarde son personnel. La Hongrie, en 1890, a été la première à utiliser cette indication; elle fut suivie par la Belgique et par la France en 1896; depuis 1901, cette organisation est entrée définitivement en vigueur dans le recensement français qui a pris des mesures pour que les bulletins des ouvriers et employés appartenant à une même entreprise soient groupés dès le moment où les mairies opèrent les premiers classements. Ce n'est pas le moment d'entrer dans le détail de ces opérations. Elles seront exposées *in extenso* dans la partie qui a trait aux recensements, mais c'est ici qu'il convient de dire quelques mots de la valeur critique du procédé.

Cette méthode est très efficace pour mettre obstacle à ce

que des entreprises soient omises ou soient comptées plusieurs fois, pour caractériser la nature de l'industrie exercée et pour empêcher que les patrons n'omettent de déclarer le personnel qu'ils occupent. Mais le point intéressant consiste dans la conformité des déclarations patronales et ouvrières. Il peut y avoir identité entre le nombre de bulletins d'ouvriers annexés au bulletin d'entreprise et le nombre d'ouvriers et d'employés déclaré par le chef d'entreprise ; dans ce cas, il n'y a évidemment aucune modification à apporter au bulletin patronal, mais les règles adoptées dans les deux autres cas par le service du recensement français aboutissent, en fait, à compter au maximum les ouvriers : si le nombre déclaré par le patron est inférieur au nombre de bulletins trouvés, on le porte, en l'augmentant, à ce dernier chiffre ; s'il lui est supérieur, on ne le réduit pas, sans doute parce qu'on suppose que les bulletins ouvriers manquants se sont égarés. Il semble que cette règle, à la considérer avec quelque attention, ne soit pas aussi satisfaisante qu'elle le paraît à première vue. Elle s'inspire évidemment de la pensée que le maître a intérêt à dissimuler le chiffre vrai de son personnel, alors que l'ouvrier n'en a aucun à déclarer qu'il travaille pour le compte de tel patron plutôt que pour le compte de tel autre. Ceci est exact, mais ce n'est pas toute la question ; l'ouvrier a aussi un intérêt à faire une déclaration : celle qu'il n'est pas en état de chômage, et plutôt que de se faire connaître comme chômeur il n'hésitera pas à inscrire sur son bulletin le nom d'un ancien patron ou de toute autre personne. Entre ces deux intérêts, il n'y a pas possibilité de se prononcer d'une façon décisive, aussi la règle rapportée plus haut a-t-elle le défaut, dans sa généralité, de trancher d'une façon systématique des cas douteux. Il paraît plus satisfaisant de s'en tenir aux déclarations émanant des ouvriers, dans tous les cas, ou, ce qui vaudrait mieux, de prendre texte de ces déclarations pour réclamer au chef d'entreprise des éclaircissements complémentaires.

La statistique des grèves présente aussi des données qui, empruntées à une source unique, seraient fort sujettes à caution. Après un conflit qui a excité toutes sortes de passions, il est évident que le récit qu'en fera chacune des parties en cause différera sensiblement. Réclamer uniquement aux patrons les renseignements concernant les grèves qui viennent de se produire dans leurs établissements, c'est admettre à l'avance que la statistique sera unilatérale et tendancieuse. Aussi, s'est-on déterminé à consulter en outre sur les circonstances, les causes et les résultats des grèves, les associations ouvrières. Tel est le système suivi en Italie, où l'on réclame en outre aux Préfets des renseignements analogues. Les déclarations ouvrières contrôlent le patronat, et vice-versa; le document dressé par l'autorité administrative sert d'épreuve et indique en quel sens la balance doit pencher. On peut y ajouter les extraits de journaux qui ont relaté l'incident. De cette façon, « statistiquer » une grève devient un véritable travail de critique exigeant beaucoup de patience unie à un grand bon sens et à une parfaite équité.

Au lieu de comparer des données venant de source différente, mais se rapportant à une même recherche, on peut encore se livrer à des aperçus critiques basés sur la confrontation de données portant sur le même sujet, mais recueillies à des moments différents par des autorités diverses, chacune dans un but particulier. Le degré d'instruction, par exemple, peut, dans une population donnée, s'apprécier par trois moyens différents : 1° en posant directement la question : « savez-vous lire et écrire » dans le recensement général de la population; 2° en classant les conscrits d'après leur degré d'instruction au moment où ils sont versés au corps; 3° en relevant le nombre d'époux qui n'ont pu signer l'acte de mariage, soit que leur ignorance résulte de la croix apposée en lieu et place de leur signature, soit qu'elle soit constatée par leur déclaration.

Indépendamment de l'intérêt que présente la comparai-

son de ces données à un point de vue général, on peut en tirer des conclusions quant à l'exactitude présumée de chaque relevé : les concordances constituent un indice favorable, les discordances servent à faire naître un doute et à provoquer une information complémentaire.

115. B. — *Crainte de désavantages d'ordre moral.* — L'intérêt, qui incite à faire des déclarations inexactes, peut ne pas être un intérêt matériel. Certaines questions des recensements peuvent provoquer des susceptibilités d'ordre moral. Les demandes concernant les cultes nous paraissent rentrer dans cette catégorie. Parmi les recensements effectués en Europe, les uns s'abstiennent de poser la question, les autres formulent la demande avec plus ou moins de détails. En Angleterre, en France, en Belgique, en Portugal, il n'y a pas de questions posées en ce qui concerne la religion des recensés ; au contraire, ce point fait partie des investigations du recensement en Allemagne, en Autriche, en Espagne, en Norvège, en Russie, en Suède, dans les Pays-Bas, etc. Il peut arriver que beaucoup de déclarations faites sur ce point soient suspectes. On ne peut non plus s'attendre à une grande sincérité de la part des recensés quand le gouvernement s'enquiert des tares physiques ou morales : le dénombrement des idiots, des fous, des sourds-muets, habitant avec les membres de leur ménage, ne sera jamais complet, car les familles cachent aussi longtemps qu'elles le peuvent les infirmités dont leurs membres sont atteints. On peut aller plus loin et altérer la vérité, afin de retarder l'accomplissement d'un devoir civique pénible ou onéreux. M. Benini a signalé que si, dans l'Italie méridionale, le nombre des naissances mâles en décembre descend au-dessous de la normale, tandis qu'au contraire le mois de janvier est au-dessus de la moyenne, la raison en est à la coutume des parents de différer de quelques jours les déclarations des naissances mâles survenues à la fin de

décembre, afin de retarder d'un an l'appel sous les drapeaux.

116. C. — *Paresse, négligence, mauvaise volonté des recensés et des agents recenseurs.* — Pour pouvoir exécuter une statistique dans de bonnes conditions, il faut pouvoir compter sur la conscience, le soin et la fidélité des agents qu'on emploie ; dans une certaine mesure, ces qualités doivent exister chez les personnes qu'on interroge ; elles doivent se retrouver au plus haut degré chez tous ceux qui dirigent le travail ou collaborent à la direction :

a) Pour obtenir des résultats satisfaisants, le chef d'un bureau de statistique doit absolument pouvoir compter sur le bon vouloir des agents recenseurs et des commis occupés au travail de dépouillement. Une des conditions requises pour que ce bon vouloir existe et se manifeste est une équitable rémunération accordée aux uns comme aux autres. Contraindre des agents inférieurs à exécuter une besogne qu'ils font à contre-cœur ou dont les conditions leur déplaisent est une détestable politique. Les économies qu'on pense réaliser par ces moyens finissent par occasionner une dépense double. L'agent recenseur, notamment, a un rôle très important à remplir. La vérification des bulletins ou formulaires lors de la reprise à domicile de ces documents, lui incombe. C'est à lui également qu'il faut avoir recours s'il s'agit d'obtenir des renseignements complémentaires ou de rectifier des erreurs. Quelles que soient les précautions prises en vue du choix de ces agents et si complètes que soient les instructions qui leur sont données, la vérification de leur travail est d'une nécessité absolue. Le plus élémentaire bon sens commande de s'assurer de la façon dont les instructions ont été exécutées. Les erreurs peuvent être tantôt accidentelles et isolées lorsqu'elles sont dues à l'interprétation erronée d'une seule personne qui, par exemple, se croit, à tort, dans l'obligation de répondre au bulletin alors qu'elle ne rentre pas dans les catégories vi-

sées par la statistique. On peut aussi classer parmi les erreurs accidentelles, l'interprétation trop large donnée au relevé par un agent recenseur isolé. Mais à côté de ces erreurs accidentelles, les administrations locales et les agents recenseurs en commettent d'autres qui ont un caractère systématique et général. Ainsi, dans un recensement industriel on constata que certaines administrations communales avaient délivré des bulletins à des personnes exerçant des professions libérales ou agricoles, à des ménagères, à des fonctionnaires et même à des personnes sans profession. Par contre, dans quelques communes, des omissions singulières avaient été commises : par exemple, on n'avait recensé que les patrons sans remettre de bulletin individuel aux ouvriers, ou l'on s'était borné à remettre des formulaires aux industriels et aux commerçants dont l'exploitation semblait présenter une certaine importance. Dans une petite ville, les agents recenseurs avaient laissé en dehors du recensement toutes les personnes du sexe féminin (1). Après des exemples tels que ceux-là, personne ne pourra plus douter de l'utilité d'une revision des plus attentives.

b) Les recensés, de leur côté, peuvent apporter leur contingent de mauvaise volonté et de négligence dans l'ensemble des facteurs défavorables à l'exactitude des résultats. Sans être altérées par l'appât du gain ou par la crainte d'un désavantage, les réponses des recensés peuvent se trouver singulièrement diminuées de valeur à raison de la négligence avec laquelle elles sont formulées. Si l'on a des raisons de soupçonner que ce défaut est général, on devra se montrer d'autant plus rigoureux dans l'examen critique du bulletin.

117. Enfin, l'éditeur même du travail a sa part de responsabilité. On peut appliquer à un travail statistique les

(1) Belgique : Recensement de l'industrie et du commerce au 31 décembre 1910. Vol. I. *Exposé des méthodes*, ch. III, paragr. 2 et 3. Bruxelles, 1913.

méthodes critiques qui s'appliquent à toute œuvre littéraire et se demander dans quel esprit le travail a été fait. A-t-on voulu exposer la vérité, et rien qu'elle? N'a-t-on pas été préoccupé d'accumuler des matériaux dans le but de produire de l'effet, d'en imposer par la masse des données? Ce sont des questions très délicates, aussi le critique fera bien de n'avancer que pas à pas sur ce terrain dangereux. Aujourd'hui, on attend de l'éditeur d'un travail qu'il fasse l'exposé complet de ses sources et de sa méthode. Cette habitude est entrée dans les mœurs en ce qui concerne plusieurs branches scientifiques, les sciences historiques, par exemple. Il existe dans chaque pays des ouvrages et des revues consacrés aux questions de méthode historique. Bien qu'il y ait une tendance sérieuse et générale en statistique à suivre cet exemple, on ne peut dire encore que plus rien n'est à souhaiter. Beaucoup de publications officielles se montrent encore d'un laconisme par trop grand dans la description des procédés suivis dans les travaux qu'elles présentent.

Est-il besoin de dire qu'en aucun cas le statisticien n'a le droit de modifier les chiffres obtenus à l'aide du relevé? S'ils ne sont pas complets, il se bornera à signaler les lacunes et à en exposer brièvement les motifs en note au bas de la page. Dans le cas où, à l'aide de méthodes spéciales, il aurait réussi à faire disparaître certaines lacunes, il doit le faire observer et faire connaître les procédés employés. En cas de discordance de certains chiffres avec certains autres, l'erreur doit être recherchée jusqu'à ce qu'elle soit découverte et corrigée. Toute discordance n'est pas le résultat d'une erreur : elle peut aussi provenir de l'emploi de méthodes différentes de comptage; ces cas seront soigneusement exposés dans l'exposé des méthodes et l'explication rappelée en note. On n'a jamais le droit de modifier sciemment un chiffre afin d'obtenir une concordance trompeuse. Nous aimons à croire que ces règles, les plus élémentaires de la probité scientifique, ont été toujours et partout res-

pectées par les statisticiens. Aujourd'hui personne n'admettrait que des manquements à ces principes puissent être commis et nous croyons pouvoir dire qu'effectivement il ne s'en commet point dans une mesure appréciable.

CHAPITRE III

Critique d'exactitude

118. Les causes d'erreur relevées par la critique de sincérité dérivent toutes des mobiles que les anciens traités comprennent sous le nom d' « animus observandi », soit dans le chef de l'organe de l'observation statistique, soit dans le chef de celui qui se prête à l'observation. Par leur nature, qui est d'essence psychologique, elles sont susceptibles d'une analyse systématique relativement simple : il suffit d'énumérer et de classer les raisons qui peuvent agir pour diminuer la sincérité des réponses des recensés, comme celles qui ont une action sur la diligence des organes du relevé. Au contraire les erreurs que la critique d'exactitude a pour objet de signaler sont accidentelles. Il est donc impossible d'en dresser d'avance une liste complète et il faut vérifier, dans chaque cas, les causes qui ont pu vicier le document.

L'examen auquel donne lieu cette recherche peut se diviser en deux parties distinctes : *a*) vérification interne; *b*) vérification externe.

La vérification interne elle-même se subdivise en trois parties : *a*) recherches des lacunes possibles du relevé; *b*) recherche des multiples emplois, ou recensement double (multiple, en certains cas) des mêmes unités; *c*) recherche des erreurs accidentelles, contradictions et autres défauts des réponses.

La vérification externe consiste à s'assurer que tous les détails matériels de l'exécution, calculs de tout genre, im-

pression des résultats, etc., sont conformes aux résultats obtenus.

La critique d'exactitude est l'une des parties les plus pénibles du travail statistique. Elle suppose chez celui qui s'y livre une attention extrême, une patience à toute épreuve, une constance que ne rebute aucune besogne aussi fastidieuse qu'on puisse l'imaginer, une probité au-dessus de tout soupçon. Ces qualités si rares doivent exister non seulement dans la direction, mais aussi parmi les agents subalternes. Le public apprécie trop peu la valeur morale et intellectuelle requise pour les travaux de statistique; peut-être le lecteur, après avoir parcouru les pages qui suivent, s'en formera-t-il une idée plus favorable et surtout plus juste.

I. — Vérification interne.

119. a) *Recherche des lacunes de la statistique.* — Aucune statistique quelconque, — nous l'avons fait observer au début de cet exposé, — ne peut prétendre à une exactitude absolue : celle-ci serait d'ailleurs hors de proportion avec les résultats à atteindre et l'usage qui est fait de ces résultats. Mais entre l'exactitude absolue et l'erreur véritable, il y a un degré intermédiaire, beaucoup plus rapproché du premier terme que du second, qu'il faut absolument atteindre. Il est impossible de tracer des règles théoriques indiquant la marge tolérable, parce que cette limite est essentiellement variable d'après les sujets traités et d'après le nombre d'unités relevées. Le bon sens indique suffisamment qu'une erreur de 10 unités sur le total est très différente quand le chiffre exact est 100 et quand il est 1,000. L'importance de l'erreur, et par conséquent la tolérance admise dépend aussi de la nature de l'unité; il y a des unités égales comme poids statistiques, — les individus compris dans un recensement de la population — et il en est d'autres dont le poids varie prodigieusement — l'unité « entreprise » dans un recensement industriel, par exemple.

Dans le premier cas, une unité en vaut une autre, l'absence de l'une n'exerce pas une action plus marquée sur le résultat final que l'absence de l'autre ; au contraire, dans le second cas, certaines unités en valent cent et mille autres. L'omission d'un grand établissement industriel diminuerait le total recensé d'un millier (ou plus) d'ouvriers, d'employés, de chevaux-vapeur, alors que celle de l'entreprise d'un charron de village ne ferait que réduire d'une unité ou deux le nombre des exploitants et celui des membres de la famille des exploitants. De plus, si un établissement industriel important se trouve omis dans le relevé, il peut arriver que ce soit dans une catégorie qui ne compte que très peu d'entreprises ; la lacune a alors une gravité proportionnelle à la rareté des unités comprises dans cette catégorie. On voit par là que la recherche des unités omises n'a pas partout une égale importance et que les résultats de cette recherche n'ont pas dans tous les cas la même valeur. L'attention et les efforts déployés seront naturellement en proportion de ceci.

Comment procéder pour constater ces lacunes ? Que faut-il faire pour les faire disparaître ?

Nous répondons à la première question en énumérant quelques-uns des moyens de vérification qu'il est possible d'employer. Cette énumération n'est pas complète, elle ne peut l'être, chacun le comprend. Chaque difficulté trouve sa solution si elle est étudiée par un homme intelligent et surtout consciencieux, mais il y a autant de solutions possibles que de cas différents ; on est donc en dehors des limites de l'énumération et les procédés indiqués ci-après ne sont qu'exemplatifs.

Il y a d'abord une sorte d'analyse logique qui fera reconnaître si dans les résultats du relevé, il ne se rencontre rien d'invraisemblable. Dans un dénombrement de la population, il paraîtrait singulier que le nombre d'hommes fût, dans une localité, deux fois plus grand que le nombre de femmes, ou vice-versa, à moins que dans cette localité il y

ait une raison qui explique cette anomalie : une forte garnison militaire, un grand collège de garçons, une maison de correction pour jeunes gens, etc. En dehors de circonstances de l'espèce, un écart très marqué dans la proportion des deux sexes est un indice qu'il ne faut pas négliger dans la lutte engagée avec l'erreur. S'agit-il d'un recensement industriel? Ici les contrôles basés sur la simple vraisemblance se multiplient. Ainsi, la présence d'ouvriers masculins isolés au milieu d'une très nombreuse population ouvrière féminine, dans des métiers de femmes, attirera l'attention et est de nature à justifier quelques recherches complémentaires. Ce n'est pas à dire que l'inscription d'ouvriers masculins soit toujours, dans un cas de l'espèce, le résultat d'une erreur; on ne peut oublier que le vrai n'est pas toujours vraisemblable, mais elle doit susciter l'attention et provoquer un supplément d'information. Nous trouvons des ouvriers mâles exerçant la profession de dentellière; des mains masculines paraissent inhabiles à manier les fuseaux; cependant, vérification faite, le relevé n'était pas erroné; c'est que la technique dentellière est si profondément ancrée dans certains centres de population qu'elle s'apprend par imitation et que des chômeurs ou des invalides s'y appliquent pour gagner un faible salaire.

120. L'exposé des différents points de contrôle utilisables excéderait les limites d'un ouvrage général tel que celui-ci. Il faut se borner à indiquer le principe et à en faire une application. La sagacité et aussi l'ingéniosité scientifique de celui qui se livre à ce travail de revision sauront tirer des résultats nombreux et variés de l'analyse logique du document.

Au nombre des moyens employés le plus fréquemment, signalons la comparaison établie entre les données nouvelles et les résultats d'un autre relevé portant sur un même objet, mais exécuté antérieurement et utilisant des informations provenant d'une autre source. Les résul-

tats obtenus à l'aide d'un dénombrement général sont ainsi souvent comparés à ceux indiqués par un relevé ne portant que sur certaines parties de l'ensemble. Ce procédé suppose que la statistique spéciale présente de plus grandes garanties d'exactitude que la statistique générale. Ceci est parfaitement d'accord avec la logique. Pour les statistiques spéciales, on a recours habituellement à des agents plus qualifiés que les simples recenseurs; souvent aussi leurs fonctions habituelles les mettent en rapport quotidien avec les personnes ou les entreprises qu'ils sont, par occasion, chargés de recenser. Il se comprend aisément que les ingénieurs du corps des mines soient à même d'opérer un recensement plus exact des charbonnages que celui obtenu à l'aide d'un dénombrement portant sur toutes les industries du pays, puisque leurs fonctions administratives les mettent en contact journalier avec les entreprises houillères situées dans leur ressort. Aussi, est-il tout indiqué que le statisticien chargé d'un dénombrement général compare les données qu'il a obtenues avec celles fournies par la statistique spéciale des charbonnages. C'est une garantie excellente à l'égard des lacunes qui peuvent si facilement se produire dans un relevé exécuté par des agents peu qualifiés, disséminés sur tout le territoire d'un pays. De même pour d'autres parties du relevé : les données résultant des documents groupés par les commis des accises, celles réunies par les inspecteurs du travail, celles provenant d'enquêtes spéciales seront mises à profit pour évaluer et corriger les erreurs toujours à craindre dans les dénombrements généraux. De nombreuses sources peuvent être ainsi consultées pour peu qu'on y mette de soin et d'ingéniosité. Nous supposons évidemment que les relevés à comparer portent sur les mêmes unités et aient été exécutées à peu près à la même époque.

Un autre système de vérification consiste dans la comparaison des données numériques avec celles obtenues an-

térieurement. Il ne s'agit plus ici que de recueillir des indices généraux au sujet des lacunes possibles du relevé. Les modifications trop profondes, les changements trop brusques feront l'objet d'une étude spéciale. Sans qu'elles indiquent toujours des erreurs, elles auront du moins l'avantage d'attirer l'attention. Il faut que l'histoire économique des années qui séparent les deux relevés fournisse la justification théorique des modifications plus ou moins radicales. Si cette justification fait défaut, ou si elle est manifestement incomplète, il y aura lieu d'approfondir les recherches. Admettons qu'une catégorie d'entreprises — celle du vêtement par exemple — ne soit plus représentée que par un nombre beaucoup moindre d'ateliers que par le passé, il y a lieu de se demander si les agents du dénombrement le plus récent ont bien distribué des bulletins à tous les tailleurs et couturières. Pour nous en rendre compte, nous commençons par délimiter le champ des recherches, puis nous verrons si le nombre des ouvriers des entreprises les plus importantes n'a pas augmenté dans une mesure telle que le fait explique la disparition d'une partie des petits ateliers ; nous aurons ensuite à nous demander s'il ne s'est pas produit une modification générale des procédés de travail (travail mécanique au lieu du travail manuel) ; nous irons même jusqu'à consulter les chiffres du commerce international afin de nous assurer si les importations de vêtements n'ont pas augmenté dans une mesure qui explique la suppression d'ateliers moins bien outillés que ceux des concurrents de l'étranger, ou si l'explication ne doit pas être cherchée dans la réduction des exportations. Les procédés graphiques, dont il sera parlé plus loin, jouent également un rôle intéressant dans ce genre de confrontations. Enfin, par la comparaison des données obtenues pour les différentes localités, on s'assurera si l'erreur est généralisée ou seulement propre à certaines divisions territoriales. Si une erreur est constatée, les mesures voulues seront prises pour la faire

disparaître, notamment en prescrivant des recherches nouvelles accompagnées d'instructions complémentaires détaillées.

121. b) *Recherche des multiples emplois.* — On désigne sous le nom de « double ou multiple emploi » le fait que certaines unités statistiques ont été comptées plusieurs fois au cours du relevé. Certaines statistiques échappent presque entièrement au danger des doubles emplois, d'autres y sont fort exposées. C'est encore à la statistique de l'industrie — l'une des plus difficiles — que nous emprunterons nos exemples. Il n'y a guère de danger qu'une entreprise de minime importance soit comptée plusieurs fois; le risque se trouve plutôt du côté opposé, c'est-à-dire qu'elle ne soit pas recensée du tout ou qu'on lui attribue une qualification économique inexacte, mais pour les établissements importants, notamment les sociétés anonymes, les comptages multiples sont à redouter. Il est assez fréquent que le chef d'un établissement de l'espèce n'habite pas la commune où sont situées les usines. Dans ce cas, s'il s'agit d'un dénombrement de personnes, il est compté dans la localité qu'il habite et est invité à donner tous les renseignements touchant son entreprise. Des administrateurs de la société peuvent se croire obligés à fournir les mêmes données et il en est encore ainsi du directeur et des chefs de service. Il arrive que la même entreprise soit recensée trois et quatre fois. On obvie au danger des comptages multiples en groupant tous les établissements par industrie et d'après la commune où ils sont situés. Les bulletins se rapportant à une même entreprise se trouvant ainsi juxtaposés, il est facile de supprimer ceux qui font double emploi en conservant seulement celui dressé par la personne appelée réellement à répondre. On trouvera dans les exposés des méthodes suivies dans les recensements, de nombreux exemples de comptages multiples de ce genre. L'un de ceux qui ont été signalés par un

recensement récent (1) concerne les entreprises commerciales : le cumul, par une seule personne, de plusieurs négoces de minime importance ne justifie pas la création d'autant de bulletins séparés, correspondant à des entreprises distinctes, non plus que l'exercice d'une profession pour le compte de plusieurs employeurs ne doit pas avoir pour conséquence le comptage, trois ou quatre fois répété, du même individu. L'organisation même du dépouillement — en admettant que les bulletins originaux soient transmis à l'organe central — permet le plus souvent de faire disparaître de pareilles causes d'erreur.

Un autre exemple de multiple emploi peut être trouvé dans les statistiques agricoles à propos de l'étendue cultivée. Les indications des diverses cultures sont fournies par les cultivateurs eux-mêmes et consignées au bulletin d'après leurs indications. De nombreuses erreurs sont commises. Les contenances des diverses cultures sont indiquées le plus souvent de mémoire et en chiffres ronds. Les cultivateurs pourraient aisément fournir des indications exactes en recourant aux actes de vente et de location, ou simplement aux feuilles de contribution; mais ils s'en abstiennent en général; de plus, les terrains cultivés sont souvent situés sur plusieurs communes et les délimitations des communes ne sont pas toujours exactement connues des paysans qui, au lieu de ces divisions administratives, ont recours à des appellations locales. De tout ceci, il résulte que les contenances indiquées et la répartition des terrains entre les diverses communes sont souvent fautives. Il arrive que l'addition de toutes les parcelles déclarées pour une commune donne un total supérieur à la superficie entière de la localité. C'est pour empêcher une erreur de ce genre que les services de la statistique agricole procèdent à une revision minutieuse de ces données.

(1) Belgique : Recensement de l'industrie et du commerce au 31 décembre 1910. Vol. I, *Exposé des méthodes*, p. LXVII.

122. c) *Recherche des erreurs involontaires, contradictions, etc.* — Elle nécessite un examen minutieux de chaque bulletin. D'après les instructions données aux agents recenseurs, ceux-ci ont déjà dû vérifier si chaque recensé a répondu à toutes les questions qui le concernent. Rien ne nous dit que ces instructions ont été suivies partout. Une revision effectuée sur le matériel même s'impose avant de passer à d'autres phases du travail statistique. La difficulté de cette recherche est allégée dans une certaine mesure quand l'examen porte successivement sur les points suivants : 1° *l'auteur du document a-t-il répondu à la question?* Les lacunes sont extrêmement fréquentes dans un matériel étendu. On demande d'indiquer la profession principale et séparément la profession secondaire : cette dernière est souvent omise. On s'informe des spécialités professionnelles : le recensé ne répond pas à la question, ou il y répond en termes vagues qui n'apportent aucun éclaircissement. Des gérants d'établissements commerciaux négligent de mentionner le nom de la personne ou de la société pour le compte de laquelle ils gèrent cette subdivision d'entreprise. Dans les registres des sociétés d'assurance contre les accidents, on omet parfois l'indication des frais de maladie ou la durée de l'incapacité subie, etc. Une lacune de l'espèce doit être complétée à l'aide des renseignements recueillis au cours de recherches complémentaires.

123. 2° *L'auteur du document a-t-il compris la question?* Malgré tous les soins apportés à la rédaction du bulletin, de fréquentes erreurs proviennent de ce fait. Les unes sont dues à l'ignorance, les autres au manque de réflexion, d'autres encore au fait que la terminologie du document n'est pas comprise. Lorsqu'on s'enquiert auprès d'un ouvrier s'il rentre dans la catégorie des ouvriers à domicile, a-t-il saisi exactement ce qu'on entend par cette expression : « ouvrier à domicile »? Le bulletin français, au lieu de cette expression, emploie celle

d'ouvrier à façon travaillant à domicile : le saisit-on toujours bien ? Les recensements professionnels distinguent entre « profession principale » et « profession accessoire ». Quand une occupation supplémentaire devient-elle une « profession accessoire » ? Et de deux professions exercées par la même personne, laquelle doit être qualifiée d'accessoire ? Les instructions précisent, autant qu'il se peut, ces divers points ; seulement, il est permis de se demander si elles ont été lues et surtout comprises.

124. Pour éviter ces erreurs ou pour les reconnaître, si, malgré tout, elles se produisent, on a recours à des « questions de contrôle ». Ainsi, après avoir réservé une partie du bulletin individuel à l'inscription de la profession et de la qualité d'ouvrier, on ajoute : « pour qui travaillez-vous ? » Les petits patrons qui auraient utilisé mal à propos cette partie du bulletin, s'empresseront de répondre : « pour les personnes de la localité ou des alentours », mais sans désigner nominativement un patron qui les aurait fait travailler. Cette circonstance suffit, le plus souvent, à les faire classer dans la catégorie des chefs d'entreprise. D'autre part, on a délimité dans la plupart des pays la sphère économique de l'industrie à domicile ; rien de plus simple que d'éliminer de la catégorie des ouvriers à domicile une série de professions qui, de toute évidence, n'appartiennent pas à la production décentralisée. Dans un recensement commercial, il y a lieu de s'informer si une exploitation industrielle n'est pas jointe à l'exercice du commerce : ainsi, après s'être informé du commerce, on posera la question suivante : « fabrique-t-on chez vous, ou faites-vous fabriquer au dehors par des ouvriers travaillant pour votre compte, certaines marchandises que vous vendez ? » et l'on ajoute immédiatement la question de contrôle : « si oui, avez-vous rempli le bulletin relatif à cette fabrication ? » L'usage des questions de contrôle est fort étendu. Il suffira de ces exemples pour donner une idée de leur nature.

L'étendue des opérations de revision ressort des chiffres suivants : en Belgique, en 1910, le service du recensement de l'industrie et du commerce a annulé 95,000 bulletins et renvoyé pour correction aux agents recenseurs 383,000 bulletins sur un total de 2,130,000. Cette revision eût été impossible si les bulletins n'avaient pas été communiqués dans leur forme originale, à l'organe central du recensement. Ce fait, mieux que tout autre, montre l'importance de la centralisation complète de tous les travaux de critique et de dépouillement.

125. Les procédés de revision dépendent en premier lieu de l'organisation administrative du service central. Ils sont donc susceptibles de prendre diverses formes. Pour ne pas se borner à cette simple constatation, indiquons quelques-uns de ces procédés, sans prétendre le moins du monde qu'il n'en existe pas d'autres. En premier lieu, beaucoup d'erreurs relevées par la critique d'exactitude sont de nature à être corrigées immédiatement, à la simple lecture du bulletin. Si les erreurs ont échappé à l'agent recenseur, elles seront relevées par l'organe central. Telles sont celles qui résultent à l'évidence des réponses données aux questions de contrôle. Ces corrections sont appelées « corrections d'office ».

Pour les recherches complémentaires, ayant pour but de faire disparaître les lacunes des bulletins, il faut s'adresser aux agents recenseurs eux-mêmes. Leur mission n'a pas été remplie par eux avec tout le soin désirable ; il leur incombe de faire compléter les bulletins en les représentant de nouveau aux recensés, à leur domicile. C'est seulement après vérification des réponses complémentaires qu'il sera opportun de liquider les indemnités attribuées aux agents recenseurs.

Les difficultés sont-elles plus sérieuses et faut-il procéder à une vérification plus approfondie? Il n'est guère à recommander de s'adresser aux agents recenseurs eux-mêmes. L'organe central a le choix entre deux moyens :

s'adresser, par correspondance, aux recensés, ou aux organes du relevé automatique s'il s'agit d'une opération de l'espèce. Actuellement, le développement du réseau téléphonique offre des facilités qui n'existaient pas il y a quelques années. L'organe central peut aussi envoyer en mission des délégués spéciaux lorsqu'il s'agit de données délicates à recueillir d'après un plan uniforme. Le vaste recensement des salaires compris dans le recensement belge de 1896 a été exécuté presque tout entier d'après cette méthode.

II. — Vérification externe.

126. De la vérification externe il y a peu de chose à dire; il suffit presque de la définir. Elle consiste à s'assurer de l'exactitude des calculs, des résultats partiels et globaux qui représentent l'aspect quantitatif des phénomènes.

Considérés avec attention, les chiffres peuvent mettre sur la piste de quelque erreur restée jusque-là inaperçue; l'examen auquel on les soumet est une partie de la vérification externe, celle qui touche le plus près à la critique. Si l'on donne un peu d'extension au terme « critique », on comprendra également sous cette dénomination toutes les investigations qui, par leur nature, peuvent conduire à des résultats plus exacts. C'est à ce titre que les vérifications portant sur les calculs et sur l'exactitude de la reproduction typographique appartiennent à la critique externe.

Encore qu'il ne s'agisse que d'un travail matériel, cette besogne n'est pas indigne d'occuper le statisticien. Tout le travail antérieur vient comme se cristalliser dans un nombre. N'est-il pas, dès lors, d'une grande importance que ce nombre soit correctement calculé et exactement reproduit? Malgré l'indulgence qu'il convient d'apporter dans l'appréciation des fautes matérielles quand la méthode générale est bonne, on ne peut se défendre d'un malaise réel lorsque les résultats numériques sont entachés de quelque inexactitude. Ce sont là des verrues très déplaisantes sur un beau

visage. On ne peut, sans doute, les éviter tout à fait; au moins, faut-il les redouter et agir en conséquence.

CHAPITRE IV

Précision des résultats

127. L'école mathématique groupe sous l'appellation de « accuracy » (précision), l'ensemble des procédés à l'aide desquel l'exactitude d'un relevé peut être exprimée, quand on le compare à un autre. Elle a recours aux mathématiques pour calculer l'effet que plusieurs erreurs peuvent produire, lorsqu'elles sont additionnées, multipliées ou divisées entre elles. Elle joint généralement à cet exposé les procédés de réduction des nombres en chiffres ronds, — le but de cette réduction étant de ne pas donner aux résultats du relevé une précision plus grande que celle qui peut leur être légitimement attribuée.

Cette matière peut être exposée à la suite de la critique statistique, dont elle forme comme la conclusion : c'est pour cette raison que nous la comprenons dans la section II de cet ouvrage. M. Bowley définit ainsi la règle fondamentale de la recherche de la précision, autrement dit le calcul de l'importance des erreurs : « l'erreur, dans une observation approchée (an estimate) est le rapport de la différence entre cette observation et la valeur véritable à l'observation approchée (1) ». M. Bowley continue : Représentons par u la mesure d'une certaine quantité, sachant que la valeur véritable de cette quantité est u^1 ; l'erreur dans l'estimation approchée — à laquelle nous donnerons la désignation de e, est donc :

$$e = \frac{u^1 - u}{u} \qquad (4)$$

et $u^1 = u\,(1 + e)$

La réciproque de $e \left(\frac{1}{e}\right)$ est une mesure appropriée de la précision d'une observation approchée, et cette mesure de-

(1) BOWLEY (A.), *Elements of statistics*. London, 1902, p. 201.

vient infinie lorsque l'erreur est zéro (1). On voit immédiatement que cette notion implique l'existence d'une mesure exacte, à laquelle on peut rapporter une observation du même fait, soupçonnée de renfermer une erreur d'une importance inconnue. La logique nous avertit suffisamment d'ailleurs que, pour évaluer l'erreur d'un relevé, il est de toute nécessité qu'on puisse trouver un point de comparaison en dehors, supposé plus exact que le relevé lui-même. Nous avons donc à nous demander dans quels cas et sous quelles conditions il sera possible de déterminer ce point de comparaison.

128. Quand on veut calculer le degré de précision des observations faites dans le domaine des sciences physiques, les procédés de calcul indiqués par l'école mathématique s'appliquent sans difficulté parce que, dans ce domaine, on dispose généralement d'un modèle auquel on peut comparer les résultats qui n'en sont que des copies plus ou moins approchées (2).

Dans ce cas, rien de plus simple que le calcul de l'erreur commise dans la copie.

Mais en statistique, disposons-nous de ces mesures-types

---

(1) Pour la démonstration de cette formule, cfr. BOWLEY : « Relations between the Accuracy of an Average and that of its Constituent Parts ». (*Journal of the Roy. Statistical Society*, 1897; p. 855.)

(2) En métrologie, par exemple, on dispose de mesures de longueur ou de poids ou de contenance construites avec la dernière précision. Il existe, dans tous les pays, des mètres-étalons en platine, gradués avec un soin extrême, jusqu'à donner l'appréciation du micron; ces mesures-types conservées dans des conditions qui les mettent à l'abri de toute altération, forment une base immuable à laquelle on rapporte les instruments de mesure moins parfaits construits à l'usage du commerce. Supposez que ce mètre-étalon n'existe pas, il ne faudrait pas longtemps pour que, de dégradation en dégradation, l'unité métrique usuelle s'écarte sensiblement de la longueur théorique qu'elle est censée représenter. On aurait ainsi des mètres trop longs et, sans doute, un plus grand nombre de mètres trop courts. Les transactions en seraient profondément troublées et de nombreuses contestations ne tarderaient pas à s'élever. Elles seraient insolubles, faute d'un type auquel on pourrait comparer les diverses mesures obtenues.

auxquelles nous puissions comparer nos relevés? Avons-nous la possibilité de constater nos erreurs accidentelles et d'en déterminer le quantum?

Il est clair que s'il s'agit de phénomènes physiques observés dans des conditions douteuses, on pourra presque toujours comparer ces observations à des mensurations plus rigoureuses de façon à en déterminer le coefficient d'erreur (1).

Mais, ce n'est pas dans ce domaine que se groupent les recherches usuelles de la statistique. Elles ont trait, le plus souvent, à des phénomènes sociaux ou à des faits matériels ayant des rapports étroits avec les phénomènes sociaux.

L'état de la population est un phénomène social : état quantitatif, c'est-à-dire le relevé du nombre des citoyens d'une ville, d'une province, d'un royaume; — état qualitatif, c'est-à-dire les attributs de telle population : nombre de personnes célibataires, mariées, veuves, divorcées; nombre d'agriculteurs, de commerçants, d'industriels, de personnes sans profession. La richesse publique et privée est un phénomène social par connexité aux activités sociales dont il dérive : les activités sociales sont les facteurs de la richesse, l'estimation de celle-ci, indépendamment de l'intérêt qu'elle présente par elle-même, nous fait donc connaître l'intensité des activités sociales qui la produisent. Dans ce domaine traditionnel de la statistique, nous croyons que les méthodes de comparaison signalées en vue d'établir le degré de précision du relevé, sont souvent inapplicables ou inopérantes. C'est que les opérations du relevé, lentes, pénibles et coûteuses, ne se renouvellent pas par pur dillettantisme scientifique. Si l'on a un relevé idéal, parfait au point qu'on puisse déterminer les erreurs de tous les autres en les comparant simplement au premier, il ne viendra à l'es-

(1) Par exemple, le comparateur utilisé au Bureau international des poids et mesures, à Passy, est un instrument d'une justesse extrême à l'aide duquel on peut vérifier avec la plus grande précision les variations des étalons métriques secondaires par rapport à l'étalon-type.

prit de personne de réclamer l'exécution d'une statistique qui, par définition, serait moins exacte que la première. Et si de deux relevés coexistants il est prouvé que l'un l'emporte sur l'autre, on n'aura guère besoin de recourir à des procédés de comparaison numérique, car les résultats du second seront laissés dans l'oubli du moment qu'il sera prouvé que leur valeur est inférieure. En résumé, c'est le point de comparaison qui fait défaut dans la plupart des cas : si l'on avait un relevé parfait, on n'aurait nul besoin et nulle envie d'en exécuter un autre.

Il en va autrement dans les sciences appliquées. Là, on possède une mesure-type dont toutes les autres ne sont que des copies plus ou moins réussies, des reproductions plus ou moins altérées. Il y a un mètre-étalon ; il y a des milliers de mètres en métal, en bois, en tissu, qui servent aux besoins de la vie journalière. On peut toujours comparer ces copies à la mesure-étalon et on doit le faire pour éviter de grands inconvénients.

Aussi, les hypothèses dont M. Bowley se sert, rendent bien évidente la distinction ci-dessus. Supposons, dit-il, qu'un relevé statistique nous donne ce résultat que le salaire hebdomadaire d'ouvriers agricoles soit de 13 s. et que nous sachions *qu'en réalité* le salaire de ces ouvriers agricoles est de 14 s., nous avons :

$$\frac{14-13}{13} = \frac{1}{13} \text{ ou } 7,7\ \%,$$

quotité représentant l'erreur du relevé fautif comparé au relevé exact (1). Mais pour comparer ces deux données il est nécessaire que nous sachions quelle est la réalité opposée à l'approximation ; telle est précisément toute la question. Comment connaître cette réalité, comment arriver à cette notion sinon parfaite, du moins plus exacte que l'estimation jugée par nous inférieure en précision ?

(1) Bowley (A.-L.), *Eléments of statistics.* London, 2e édition, 1902, p. 201.

129. Ne nous attendons pas à la rencontrer souvent, mais enfin elle existe dans certains cas, et si peu nombreux qu'ils soient ils suffisent à justifier théoriquement l'application du calcul aux phénomènes réunissant les conditions nécessaires. En voici un exemple basé sur la méthode générale exprimée dans la définition de l'erreur donnée par M. Bowley, et que nous avons reproduite plus haut.

Les dénombrements de personnes sont rarement comparables à d'autres relevés du même genre, portant sur les mêmes unités, effectués au même moment, deux relevés simultanés n'étant exécutés que dans des cas exceptionnels. Cependant, ce cas général comporte parfois l'une ou l'autre exception. En voici une : le compte rendu des opérations des chemins de fer de l'Etat belge, qui paraît chaque année, renferme la statistique des ouvriers appartenant à l'administration. En 1910, au 31 décembre, le nombre de ces ouvriers était de 55,600 (1). A la même date, avait lieu en Belgique le recensement général de l'industrie et du commerce qui, au moyen de bulletins individuels, relevait le nombre des exploitants, employés et ouvriers occupés en Belgique. Les cadres statistiques réservaient une place spéciale au dénombrement du personnel actif des chemins de fer de l'Etat; en y comprenant 598 chômeurs, évidemment portés sur les listes dressées par l'administration des chemins de fer, le relevé statistique de 1910 donne un total de 56,046 ouvriers (2).

D'après ce qui précède, en admettant comme exact le chiffre donné par l'administration des chemins de fer de l'Etat, on a la fraction :

$$\frac{55{,}600 - 56{,}046}{56{,}046} = \frac{446}{56{,}046} \text{ ou } \frac{223}{28{,}023} \text{ ou } 0{,}0079.$$

---

(1) Chemins de fer de l'Etat belge. Compte rendu des opérations pendant l'année 1910. Rapport présenté aux Chambres législatives : Chambre des Représentants, n° 177. Bruxelles 1911, A, p. 8.

(2) Ministère de l'Industrie et du Travail. Office du Travail. Recensement de l'industrie et du commerce, recensement professionnel; t. II. Bruxelles, 1913, p. 1414, col. 13 et 19.

La supposition que le chiffre donné par l'administration des chemins de fer de l'Etat représente exactement le nombre des ouvriers est une hypothèse assez gratuite : dans un comptage qui introduit une distinction aussi délicate que celle d'employés et d'ouvriers, rien de plus facile que de commettre une erreur. Aussi, serait-il plus prudent d'accorder une égale valeur aux résultats en présence et de considérer l'écart qu'ils présentent comme une simple discordance et non comme une erreur à imputer à l'un des relevés plutôt qu'à l'autre. Dans ce cas, la fraction devrait s'écrire :

$$\pm \frac{223}{28{,}023} \text{ ou } \pm 0{,}0079.$$

130. Mais de tels exemples sont rares; toute comparaison statistique suppose en effet que l'unité soit la même de part et d'autre. Cette condition n'est pas toujours réalisée. Dans deux opérations statistiques exécutées par des organismes indépendants, les définitions des unités seront rarement concordantes. La statistique du chômage, par exemple, présente un cas de l'espèce. Un relevé du chômage effectué lors d'un recensement général nous paraît suspect. La pensée nous vient naturellement de prendre pour point de comparaison, en lui attribuant une valeur plus grande qu'à notre dénombrement, le relevé des chômeurs dressé par un certain nombre de syndicats ouvriers. Il est raisonnable de penser que les listes de chômeurs dressées par ces syndicats sont plus exactes que toute autre, attendu que des indemnités journalières sont accordées aux syndiqués sans travail et que, dès lors, une comptabilité régulière s'impose. Mais les définitions du chômeur ne concordent pas en fait. Les syndicats ne connaissent que le chômeur indemnisé et comme le droit à l'indemnisation cesse après un certain temps, il arrive, pendant les périodes de crise industrielle, que les syndicats comptent un assez grand nombre de chômeurs qui ont épuisé leurs droits aux allocations de chômage; d'autre part, le chômage est généralement plus

étendu parmi les non-syndiqués que parmi les syndiqués. Il est donc impossible, en s'appuyant sur des documents de l'espèce, de découvrir et de mesurer l'erreur qu'un relevé général pourrait présenter.

On ne peut non plus comparer une moyenne à un chiffre obtenu directement par l'observation statistique. A une date déterminée on relève la population ouvrière industrielle d'un pays et on trouve qu'elle s'élève, dans les mines de houille, à 147,000 personnes. L'administration des mines, de son côté, publie la statistique de la population ouvrière : elle se borne à une moyenne analogue à celle du calcul du *vollarbeiter,* c'est-à-dire qu'ayant relevé le nombre de journées de travail elle divise ce nombre par le chiffre de 300 (nombre conventionnel des journées d'activité en un an) : le quotient — dans le pays dont il s'agit — donne 143,000. On ne comparera pas les deux nombres parce qu'ils sont obtenus par des méthodes différentes et ne représentent pas la même chose.

131. Il arrive assez souvent que les données relatives aux choses fournissent plusieurs points de comparaison utilisables ; les choses matérielles intéressent souvent plusieurs services chargés de la statistique, de sorte qu'on possède des données recueillies de divers côtés. Ces évaluations ne sont pas généralement concordantes, même quand elles s'appliquent à un objet parfaitement défini, tel que, par exemple, la superficie d'un pays. Voici quelques chiffres montrant ces variantes, en ce qui concerne la superficie de la Belgique :

Superficie, en hectares, de la Belgique

A. D'après la statistique des biens de mainmorte (1866), 2,945,516 hectares

B. D'après le recensement général de la population (1880), 2,945,715 hectares.

C. D'après le recensement général de l'agriculture (1880), 2,945,589 hectares.

D. D'après le recensement général de l'agriculture (1895), 2,945,557 hectares.

E. D'après le recensement général de la population (1900), 2,945,503 hectares.

F. D'après le recensement général de la population (1910), 2,945,104 hectares.

G. D'après le recensement général de l'agriculture (1910), 2,945,040 hectares.

La discordance entre les divers résultats tient à diverses causes qu'il est inutile de rechercher ici. Pour les chiffres de deux recensements, ceux de 1900 et de 1910, par exemple, la différence s'exprimera par la fraction :

$$\frac{2,945,503 - 2\,945.104}{2,945,104}$$

précédée du signe $\pm$ s'il n'y a pas de bonnes raisons pour préférer les données de l'un aux données de l'autre. Si l'on possédait un relevé présentant plus de garanties, on pourrait utilement lui comparer les données de l'un et de l'autre recensement en faisant cette fois précéder la simplification de la fraction des signes + ou — selon les cas.

Les résultats consignés plus haut sont une illustration de la règle générale d'après laquelle les différences très minimes peuvent être négligées s'il ne s'agit pas d'une description purement scientifique portant sur des mesures extrêmement précises. On peut, hormis ce cas, remplacer deux ou trois rangées de chiffres par des zéros, ou transformer ces trois rangées de chiffres en un chiffre rond à ajouter aux tranches précédentes, en la faisant précéder du signe $\pm$; ainsi, dans l'exemple donné *sub* litt. E, on écrira plus exactement le résultat relatif à l'année 1900, comme ceci : 2,945,000 $\pm$ 500.

132. Pour terminer cette matière de la précision des observations, il y a lieu d'examiner l'effet que plusieurs erreurs peuvent produire lorsqu'elles sont additionnées, multipliées ou divisées entre elles. Nous formulons ci-après les règles du calcul en suivant les formules données à cet effet par les auteurs appartenant à l'école mathématique, mais en nous limitant à l'addition et au calcul des moyennes arithmétiques simples qui sont les cas les plus fréquents.

« L'erreur dans un total d'observations, dit M. Bowley, est égale à la somme des erreurs des facteurs quand chacune

est multipliée par le rapport de la partie correspondante au total. » M. Bowley expose ainsi la formule (1) :

Soit u quantités telles que u_1, u_2, ... u_n dont le total est désigné par la lettre u, de sorte que u est égal à $u_1 + u_2 + \ldots u_n$. Appelons e la somme des erreurs de ces quantités, erreurs désignées par e_1, $e_2 \ldots . e_n$. La valeur véritable du total est $u\ (1 + e)$ et les valeurs véritables de chaque partie composant le total est $u_1\ (1 + e_1)$, $u_2\ (1 + e_2)$ .... u_n $(1 + e_n)$

donc :

$$u\ (1 + e) = u_1\ (1 + e_1) + u_2\ (1 + e_2) + u_3\ (1 + e_3) + \ldots..$$

Mais nous avons défini

$$u = u_1 + u_2 + \ldots . u_n$$

d'où, par soustraction, nous avons

$$ue = u_1\, e_1 + u_2\, e_2 + \ldots..$$

et

$$e = e_1 \times \frac{u_1}{u} + e_2 \times \frac{u_2}{u} + \ldots . \qquad (5)$$

M. Bowley donne l'exemple suivant comme application arithmétique : 2 Trades Unions indiquent respectivement 555 et 45 membres sans travail alors que les chiffres véritables sont 565 et 50, de sorte que les erreurs sont :

$$\frac{2}{111} \text{ et } \frac{1}{9}.$$

D'après la règle ci-dessus, l'erreur dans le total est :

$$\frac{2}{111} \text{ de } \frac{555}{600} + \frac{1}{9} \text{ de } \frac{45}{600} = \frac{1}{40} \text{ ou } 2\ 1/2 \text{ p. c.}$$

Si l'on fait la moyenne de plusieurs observations approchées, l'erreur de la moyenne est la somme des observations particulières quand chacune de ces erreurs est multipliée par le rapport de l'estimation correspondant à celui du

(1) Bowley (Arthur L.), *Elements of statistics*, London, 1902, p. 203.

total des estimations. La formule se déduit facilement de la précédente. Elle est ainsi donnée par M. Bowley :

Soient les estimations m_1, m_2 ... m_n de quantités affectées d'une certaine erreur e_1, e_2, ... e_n, de sorte que leur valeur véritable est $m_1 (1 + e_1)$, $m_2 (1 + e_2)$, ...

la moyenne des estimations est évidemment :

$$\frac{m_1 + m_2 + \dots . m_n}{n}$$

(n désignant le nombre de termes)

et la moyenne des valeurs exactes est :

$$\frac{m_1 (1 + e_1) + m_2 (1 + e_2) + \dots . m_n (1 + e_n)}{n}$$

L'erreur de la moyenne, d'après la règle ci-dessus, est :

$$\frac{\frac{m_1 (1 + e_1) + m_2 (1 + e_2)}{n} + \dots - \frac{m_1 + m_2}{n} + \dots}{\frac{m_1 + m_2}{n} + \dots} = \frac{e_1 m_1 + e_2 m_2}{m_1 + m_2} + \dots$$

$$= e_1 \times \frac{m_1}{\Sigma m} + e_2 \times \frac{m_2}{\Sigma m} + \dots$$

Σ désignant la somme des m.

On trouvera notamment dans Bowley des règles applicables aux multiplications, divisions, moyenne pondérée, etc. que nous omettons ici, les cas d'application étant beaucoup plus rares qu'en ce qui concerne les formules précédentes

133. Afin de ne pas citer des nombres compliqués dans les exposés des résultats statistiques, on simplifie souvent les chiffres en ne citant que les données caractéristiques. Au lieu de dire : « telle ville compte 102,021 habitants », on écrira simplement : 102,000 habitants. Les simplifications de ce genre conviennent particulièrement aux exposés rédigés en vue du grand public ou ayant un but didactique. Elles sont imposées également en vertu des principes rap-

pelés plus haut ayant trait à la précision des données numériques. On ne doit pourtant y recourir qu'avec discrétion à cause des difficultés qui résultent de leur comparaison avec les données numériques plus précises qu'on rencontre dans les tableaux de chiffres : lorsque le chercheur veut reconstituer un nombre par addition successive de ses facteurs il peut éprouver des hésitations s'il compare ces résultats aux nombres simplifiés. Pour simplifier les chiffres on suit des règles précises dont la plus essentielle est de s'arrêter au dernier chiffre dont la précision puisse être garantie. Si ce chiffre est la première décimale de chaque nombre, on se bornera à écrire le nombre entier avec la première décimale en « forçant » ce chiffre s'il dépasse 5 ou si le chiffre dont il s'agit est égal à 5 et est suivi lui-même de chiffres supérieurs à 0. Les exemples suivants précisent ce qui précède :

28.94103	se	simplifie	à la première	décimale	en 28.9
32.95140	»	»	»	»	» 33.0
19.90200	»	»	»	»	» 19.9
27.05683	»	»	»	»	» 27.1
23.04999	»	»	»	»	» 23.0
36.05001	»	»	»	»	» 36.1

Si tous les chiffres sont corrects et si leur exactitude est démontrée, il vaut mieux citer l'expression fractionnaire que d'arrondir. Une règle métallique bien construite mesure 7 m. 255 millimètres; il est préférable de citer ce nombre, dont nous supposons l'exactitude parfaite, que d'écrire 7 m. 2 ou 7 m. 26, car la mention de tous les chiffres est plus exacte que l'abréviation.

134. *Références.*

BENINI (R.), *Principii di statistica metodologica*, Torino, 1906 (critica e comparazione dei dati primitivi).

BERTILLON (J.), *Cours élémentaire de statistique administrative*, Paris, 1895, pp. 45-49.

BLOCK (M.), *Traité théorique et pratique de statistique*, deuxième édition. Paris, 1886, pp. 292-298.

BLODGETT (James H.), *Obstacles to accurate statistics* (American statistical association), 1898-99.

BOSCO (A.), *Lezioni di statistica*, Roma, 1905, pp. 293-325.

BOWLEY (Arthur L.), *Elements of statistics*, second edition, London, 1902, pp. 199-214.

ID., *An elementary manual of statistics*. London, 1910, p. 6.

DURAND, *Census methods* (American statistical association), 1908-1909, p. 609.

GABAGLIO, *Teoria generale della statistica*, Milano, 1888, vol. II, pp. 103-109.

KING (W. J.), *The elements of statistical method*, New York, 1912, p. 64.

MARCH (L.), *Statistique* (de la méthode dans les sciences, deuxième série). Paris, 1911, p. 338.

MEITZEN (A.), *History, theory and technique of statistics* (trad. angl. de Roland P. Falkner), Philadelphie, 1891, pp. 103-04, 123.

PIDGIN (Ch. F.), *Practical statistics*, Boston, 1888, pp. 28-29.

QUETELET (A.), *Lettres sur la théorie des probabilités*, Bruxelles, 1846, pp. 298-306.

Reports of the Department of Commerce and Labor. Bureau of the Census. Washington, 1912, pp. 53-54.

SEIGNOBOS (Ch.), *La méthode historique appliquée aux sciences sociales*, Paris, 1901, pp. 61-77.

VON MAYR (G.), *Statistik und Gesellschaftslehre*, Tubingen, 1914, Erster Band, pp. 100-103.

SECTION IV

Le dépouillement et la présentation des données statistiques

CHAPITRE PREMIER

Préparation du dépouillement

I. — Généralités et division de la matière.

135. Le relevé des unités place sous les yeux du statisticien un amas confus de matériaux entre lesquels il opère un choix par le moyen de la critique. Ce premier travail achevé, il reste à classer, à distinguer d'après leur nature et d'après leur grandeur les unités définitivement retenues, puis à les compter d'après les divisions établies. Cet ensemble d'opérations nouvelles constitue le dépouillement. On peut le définir : « l'opération par le moyen de laquelle s'opère la répartition des unités ou faits recueillis par le relevé, afin d'en rendre l'étude possible et même aisée ».

On peut distinguer dans le dépouillement trois opérations distinctes : 1° reconnaître les qualités individuelles de l'unité, de façon à permettre d'effectuer la répartition, d'après leur nature, des faits observés ; 2° grouper les unités de même genre dans les classes établies au préalable et selon les conditions de temps et d'espace, pour constituer les données statistiques ; 3° totaliser les résultats partiels afin d'arriver à une expression synthétique. A ces trois opérations correspondent deux phases distinctes du travail : A, la préparation et B, l'exécution du dépouillement.

Les faits statistiques, convenablement classés, forment ensemble ce qu'on appelle une « donnée statistique ». La donnée statistique n'est pas la même chose que l'unité du relevé. Cette dernière est un fait unique : un délit, une entreprise industrielle, une importation de marchandise, sont des unités statistiques. Tant de délits de telle sorte, constatés dans telle région à une époque déterminée, tant d'entreprises d'une nature spéciale, existant dans une localité, à une date précise, forment des données statistiques. Lorsque l'unité est examinée par le statisticien au moment de la première opération du dépouillement, on peut encore reconnaître ses qualités individuelles : l'entreprise industrielle dont les caractères spéciaux sont retracés par ce bulletin, placé sous mes yeux, occupe exactement 27 employés et 351 ouvriers. Pour autant que ces caractères suffisent à me renseigner au sujet de l'importance de cette entreprise, je suis fixé à cet égard autant qu'il m'est possible de l'être. Mais ces caractères individuels ne vont pas tarder à disparaître : ayant à estimer l'importance des entreprises, j'ai été amené à créer des divisions portant uniquement sur le nombre des ouvriers occupés et à établir ces divisions selon des limites arbitraires, réalisées déjà dans des dénombrements antérieurs et admises par d'autres pays. Il existe une de ces divisions pour les entreprises occupant de 200 à 499 ouvriers ; c'est là que mon entreprise de 351 ouvriers va se trouver classée avec d'autres, les unes plus importantes, les autres moins. L'agglomérat des entreprises faisant partie de cette division pour l'industrie étudiée et dans la localité dont il s'agit, forme une donnée statistique. Il y a autant de données statistiques qu'il y a de nombres, c'est-à-dire de groupements de chiffres, dans une publication. Ce qui caractérise la donnée statistique, c'est la disparition de tous les caractères individuels de l'unité et leur remplacement par des classes-limites qui enserrent, d'une manière plus ou moins étroite, une série de faits présentant des caractères semblables ou à peu près semblables.

Le processus du dépouillement consiste donc en une série de groupements, de totalisations. Il y a groupement lorsqu'on arrête les classes-limites entre lesquelles les unités seront réparties; il y a groupement encore lorsqu'on réunit les unités dans ces classes, comme on trie la correspondance dans les bureaux de poste ou comme on place des fiches dans un classeur; il y a groupement surtout lorsqu'on réunit des totaux partiels pour en faire un total général.

136. Le dépouillement est une phase extrêmement importante des travaux statistiques, mais il n'est pas, comme certains paraissent le croire, leur point d'aboutissement. Si le praticien considère sa besogne achevée avec le dépouillement, le statisticien épris de sa science considère qu'alors seulement commence la partie intéressante de sa tâche. Après avoir groupé les résultats partiels au moyen du dépouillement, il faut encore en effet les sérier, en tirer les moyennes, les soumettre à un contrôle mathématique, déterminer leurs résultats les plus probables, les soumettre à une interprétation logique. Il y a des liens si étroits entre ces diverses opérations et celles du dépouillement qu'on pourrait presque considérer les unes comme la suite logique des autres, mais il vaut mieux les étudier séparément. Ainsi, les fins administratives de la statistique sont mises à part du travail scientifique. Toute cette dernière partie conserve alors son unité et les phases dont elle se compose s'orientent vers un but unique : la connaissance scientifique des phénomènes collectifs. C'est pour ces raisons que nous avons, dans cet ouvrage, séparé en deux parties l'exposé théorique de la méthode et qu'après avoir montré les procédés matériels de recherche, nous décrivons, dans le second livre, les méthodes scientifiques de vérification et d'interprétation des résultats.

137. Il n'est pas inutile de constater que toutes les parties des renseignements recueillis ne sont pas soumises au dé-

pouillement. En effet, bon nombre de questions des bulletins ne sont posées qu'à titre de contrôle d'autres demandes, ou pour mieux assurer l'authenticité des déclarations, ou simplement dans un but administratif quelconque n'ayant aucune relation avec les données statistiques. Ces questions ont généralement leur utilité même au point de vue statistique pur, mais elles ne font point partie, comme telles, du dépouillement. Ce fut certes un grand progrès que la substitution, lors du recensement de Paris, en 1817, des états nominatifs aux états simplement numériques qui étaient la règle auparavant. En donnant seulement le nombre et non les noms des personnes comprises dans le recensement, maison par maison, « on s'expose, disait l'auteur d'un rapport de l'époque, à des erreurs considérables, dont le nombre est infini (1) ». Si utile que soit cette réforme, il n'en reste pas moins exact que les noms des recensés ne fournissent aucune matière utilisable pour le dépouillement. Ces données ne sont pas assimilables à des scories qu'on doit rejeter systématiquement, elles ont leur utilité, parfois de premier ordre, mais c'est à un autre moment, celui de la critique statistique, qu'elles ont leur rôle à remplir. Arrivé au stade du dépouillement, le statisticien n'a plus à s'en occuper.

II. — Conditions générales du dépouillement.

138. Nous avons fait remarquer que le dépouillement consiste essentiellement en une suite de groupements embrassant un nombre croissant d'objets. Du simple fait que les unités ne peuvent être présentées avec leurs caractères individuels, à cause de l'impossibilité intellectuelle de saisir les ressemblances et les différences des objets compris dans une énumération infinie, il résulte que tout dépouillement débute par une totalisation. Or, une conséquence méthodologique très importante se dégage de cette

(1) Cfr. Block (M.), *Traité de statistique*, p. 358, Paris, 1886.

constatation : c'est que le dépouillement ne peut logiquement s'écarter trop de l'état original du matériel. S'il simplifie à l'excès les caractères des unités, il nuit à la ressemblance, il altère les traits du modèle. Reprenons l'exemple donné plus haut; supposons qu'au delà de 200 ouvriers il n'y ait qu'une classe unique groupant tous les établissements industriels comptant plus de 200 travailleurs. Cette classe serait beaucoup moins significative que la série de celles établies d'habitude dans les statistiques industrielles, qui rangent encore les entreprises en trois ou quatre divisions d'après l'importance de leur personnel ouvrier (200 à 499 — 500 à 999 — 1.000 à 1,999 — 2.000 et au delà). Ces classes forment déjà des totalisations où se trouvent confondues des unités de caractère différent, quoique assez semblable encore. Plus on les raréfie, plus on augmente l'incertitude touchant la nature exacte et la grandeur des unités qui s'y trouvent renfermées. En simplifiant à l'excès, on empêche tout jugement de se former sur une base logique, l'on rend presque inutilisable le matériel primitif et l'on réduit la portée de l'utilisation de la recherche. Même avec des classes-limites bien établies, il peut être utile de faire ressortir, à l'aide de notes, les caractéristiques de certaines unités comprises dans ces classes, mais qui présentent des traits se détachant de l'ensemble.

En plus des considérations ci-dessus, il convient de remarquer que le matériel original n'est pas à la disposition des chercheurs, de telle sorte que la publication, pour produire son maximum d'utilité, doit s'écarter aussi peu que possible de la situation qu'ont fait connaître les bulletins. Comme nous ne pouvons prévoir jusqu'à quel point de détail l'un ou l'autre chercheur voudra descendre, il est bon de se montrer très prudent au cours de l'élaboration des modèles de tableaux, en ce qui concerne les éliminations à pratiquer dans le matériel original. Dans une statistique des salaires, la première exigence scientifique consiste à recueillir des salaires qui représentent effectivement ce

que gagne chaque ouvrier pendant une journée normale de travail et à rejeter les moyennes, qu'elles se rapportent à un seul individu ou à plusieurs. Mais, même avec ce principe dont l'exactitude est évidente, on peut arriver à un rendement scientifique très différent selon le mode d'organisation du dépouillement et de la présentation. En négligeant certains détails on peut nuire gravement à l'utilisation ultérieure des données statistiques. Ceci revient, en somme, à augmenter le coût relatif de la statistique : moins les renseignements donnés sont nombreux, plus s'élève leur coût relatif. Pour répondre à un nombre de problèmes convenables, une statistique des salaires doit envisager : la nature exacte de l'industrie, le mode d'exploitation des entreprises, l'importance de la localité où l'entreprise est établie, sa situation urbaine ou rurale, l'âge et le sexe des ouvriers, leur spécialité professionnelle, la nature du travail : mécanique ou à la main, etc. (1). Si l'un de ces éléments fait défaut, telle recherche intéressante peut être rendue impossible.

Cette exigence se trouve limitée par la dépense en temps et en argent que rendrait nécessaire une publication aussi étendue. Mais cette dernière considération n'enlève rien de sa force au principe précédent. La statistique doit à la recherche scientifique de présenter les résultats de ses investigations dans le plus grand détail possible. De même que, dans les sciences naturelles, l'observation *in situ* et la présentation dans les conditions de l'observation sont à la base de la méthode (2), ainsi, dans notre matière, doit-on conserver avec soin les caractères propres des unités recueillies.

(1) La statistique des salaires dans les industries textiles en Belgique au mois d'octobre 1901 (Office du Travail, Bruxelles, 1903), ne comprend pas moins de 160 tableaux analytiques au moyen desquels ces divers éléments et leurs combinaisons se trouvent présentés et étudiés.

(2) Voyez sur ce point les remarques si judicieuses de M. Gilson dans son beau livre : *Le Musée d'histoire naturelle moderne*, p. 35, Bruxelles, 1914.

Ce qui importe surtout, c'est que les données publiées répondent à un usage pratique, qu'elles renferment la solution à donner au plus grand nombre possible de questions, sans négliger, à côté du but scientifique, l'intérêt administratif. Le statisticien est parfois tenté de considérer avec quelque dédain l'un ou l'autre point dont l'intérêt lui échappe. Si la demande est formulée en vue de répondre à un but déterminé et n'est pas inspirée par une vaine curiosité, pourquoi ne pas tenter de lui donner réponse ? Pour un groupe d'administrateurs telle question, qu'on est tenté de rejeter, peut avoir autant d'intérêt qu'en présente pour le statisticien la détermination d'un coefficient. Ce qui n'est pas dépouillé et publié est pratiquement perdu (1).

139. La multiplicité des sujets auxquels touche la statistique et l'obligation de présenter les questions sous leurs différents aspects afin de permettre l'utilisation aussi complète que possible du matériel recueilli, nous avertissent suffisamment de ce que le dépouillement comportera souvent plusieurs phases successives; chacune d'elles sera relative à un groupe distinct de questions. L'ensemble des rubriques de présentation des données relatives à un groupe spécial de questions s'appelle un cadre statistique. Dans un recensement de la population, le dénombrement, par commune, de la population de droit et de fait, du territoire et des maisons et bâtiments forme un cadre statistique; il y a un lien entre la population de droit et de fait, comme il y en a un entre la population et le territoire (densité de la population) et entre le nombre de ménages et le nombre de maisons habitées. En ouvrant au hasard le recensement de la population en Belgique (1910) nous voyons que, dans l'arrondissement de Courtrai, il y a 44,597 maisons et 45,153 ménages, d'où l'on

(1) BOWLEY (*Elements of statistics*) expose sur ce point des aperçus très intéressants.

peut conclure à l'instant que, sauf exception, chaque ménage a sa maison. A Bruxelles, au contraire, il y a 58.299 ménages et seulement 20.758 maisons. La réunion, dans un même cadre, de données se rapportant à un même groupe de questions, facilite les recherches et suggère souvent des rapprochements utiles. Un autre cadre statistique sera formé en tenant compte de l'âge des habitants et de leur degré d'instruction, etc.

Pour qu'un cadre statistique soit bien composé, il doit répondre aux conditions suivantes : 1° constituer un ensemble logique ayant trait à un groupe de questions; 2e présenter les données dans un ordre rigoureux en rapprochant celles qui sont connexes; 3° ne contenir que des données homogènes, présentées d'après une base unique; c'est par une fausse économie ou par le vain désir de paraître savant en se montrant compliqué qu'on accumule les données dans un cadre unique; 4° comporter une présentation telle que les données se combinent entre elles et paraissent sous leurs différents aspects; la simple juxtaposition des données statistiques ne suffit pas à réaliser les fins propres de la présentation.

140. On peut considérer comme un cadre statistique un exposé systématique réduit à un seul élément, mais, en général, on réserve cette dénomination à des ensembles de données se groupant autour d'un même sujet. Comme exemple du premier cas, il est certain que les données suivantes forment un cadre statistique simple :

ANNÉES	Population de la France (résidente) évaluée au milieu de chaque année (1).	ANNÉES	Population de la France (résidente) évaluée au milieu de chaque année (1).
1906	39,282,000	1909	39,421,000
1907	39,279,000	1910	39,528,000
1908	39,368,000	1911 (prov.)	39,600,000

(1) Statistique générale de la France. *Annuaire statistique* 1911. Résumé rétrospectif, n° 11.

Mais ce n'est pas sous cette forme simple que se présentent d'habitude les cadres statistiques. Ils visent à grouper, dans les diverses colonnes d'un tableau, des éléments qui, rapprochés les uns des autres, fournissent aux chercheurs la réponse à diverses questions. Ainsi on rapproche l'état et le mouvement de la population comme dans le cadre suivant, emprunté à la source citée plus haut :

ANNÉES	Population résidente évaluée au milieu de chaque année.	Mariages.	NAISSANCES — Enfants déclarés vivants — Légitimes	Illégitimes	Total — masculin	féminin	Total	Mort-nés.	DÉCÈS mort-nés non compris — masculin	féminin	Total	Décès de la première année.	Excédents des naissances ou déficits.	Proportions pour 100 habitants — des nouveaux mariés	des enfants nés vivants	des décès
	Mille	Mille	Mille	Mille	Mille	Mille	Mille	Mille	Mille	Mille	Mille	Mille	Mille			
1906	39,282	306	736	71	411	396	807	37	406	374	780	116	+27	1.56	2.05	1.99
1907	39,279	314	702	71	395	378	773	37	413	379	792	101	−19	1.60	1.97	2.02
1908	39,368	316	722	70	405	387	792	37	388	356	744	101	+48	1 60	2.01	1.89
1909	39,421	308	702	68	393	377	770	36	392	363	755	91	+15	1 56	1.95	1.91
1910	39,528	308	707	67	395	379	774	36	366	337	703	86	+71	1 56	1 96	1.78

Nous examinerons plus loin le mécanisme spécial de la formation des cadres statistiques, qui peuvent revêtir une complexité croissante, mais auparavant il convient de dire quelques mots de l'aspect le plus général que présente cette question.

141. D'après Stanley Jevons (1), le professeur de Morgan (2) aurait été le premier logicien qui se soit aperçu qu'il peut exister des syllogismes dans lesquels les nombres des objets formant les différents termes du syllogisme peuvent

---

(1) Stanley (Jevons) : « On a general system of numerically definite reasonings » dans : *Pure logic and other minor works*, London 1890, pp. 173 et suiv.

(2) Cfr. A. de Morgan: *Formal logic : or, the calculus of Inference.* London, Taylor and Walton, 1847.

être exactement définis. Avant de Morgan, les logiciens se bornaient à introduire des notions quantitatives vagues, comme : *peu, beaucoup plus, moins,* etc. Peu de temps après de Morgan, le professeur Boole (1) s'est occupé de la même question sous le titre : « On statistical conditions ». La notation employée par de Morgan comme par Boole prêtait à des critiques fondées; elle a été corrigée et simplifiée par Stanley Jevons. Cette dernière notation a été adoptée, avec quelques modifications de détail, par M. U. Yule. Dans l'exemple que nous donnons ci-après, nous suivons nous-même la notation de M. Yule (2).

Comme le dit cet auteur, il est parfois nécessaire, dans un but théorique, de posséder une notation simple pour désigner les classes formées parmi les attributs d'un groupe social et pour symboliser les nombres d'observations relatives à chaque attribut. Le but d'une telle recherche, écrit Stanley Jevons, est de déterminer, d'une manière aussi exacte que possible, le nombre d'objets individuels composant des classes ou groupes d'objets, étant donnée une condition logique quelconque.

Avant d'exposer un exemple, il est nécessaire de préciser la notation employée. On désignera par A, B, C, les divers attributs observés. Tous les individus possédant l'attribut A, forment ensemble la classe A. L'attribut A est positif; on doit pouvoir désigner l'absence de cet élément; pour cela on employera les lettres de l'alphabet grec : α β γ. Donc α signifiera la classe de ceux qui ne possèdent pas A; B α, la classe de ceux qui possèdent B, mais qui se font remarquer par l'absence de A, et ainsi de suite. Ces combinaisons de lettres s'appellent des « classes symboles ». Si l'on veut exprimer le nombre d'observations que comprend

(1) Cfr. Boole : *Laws of Thought,* 1854 (ch. XIX, « On statistical conditions »).

(2) Yule (G. U.), *An introduction to the theory of statistics,* ch. I, London 1911.

chaque classe, on choisira la lettre relative à cette classe et on la mettra entre parenthèses (A), (α), (B), ( ), etc.

Soient les notations suivantes :

A = ouvriers recevant un salaire en argent;
B = ouvriers recevant un salaire en argent et la nourriture;
C = ouvriers recevant un salaire en argent, la nourriture et le logement.

Nous pouvons, à l'aide de cette notation, en introduisant les symboles négatifs, réaliser les combinaisons suivantes :

(A) = nombre d'individus possédant le caractère ou l'attribut A;
(α) = nombre d'individus ne possédant pas le caractère ou l'attribut A;
(A B) = nombre d'individus possédant à la fois le caractère ou l'attribut A + B;
(α B) = nombre d'individus possédant seulement le caractère ou l'attribut B, mais non A.
(A B C) = nombre d'individus possédant à la fois le caractère ou l'attribut A + B + C;
(α B C) = nombre d'individus possédant à la fois le caractère ou l'attribut B + C, mais non A;
(α β C) = nombre d'individus possédant le caractère ou l'attribut C, mais ni A ni B.

Les classes positives et négatives forment les classes contraires, l'une par rapport à l'autre, et leurs fréquences sont des fréquences opposées. Les classes se distinguent par ordre; le rang de leur ordre est celui du nombre d'attributs de la classe.

(A) = 1er ordre. — (A B) = 2^{e} ordre. — (A B C) = 3^{e} ordre.

Les classes du n^{e} ordre, dans le cas de n attributs, sont dénommées classes les plus élevées et leurs fréquences sont les « fréquences dernières » (*ultimate frequencies*). Lorsqu'on connaît les « fréquences dernières », toutes les autres peuvent être obtenues par simple addition.

142. Exemple : nombre des ouvriers boulangers-pâtissiers de plus 16 ans d'après la nature du salaire qu'ils reçoivent (*Recens. gén. des Ind. et des Métiers en Belgique*, octobre 1896, vol. XI, p. 317).

Nombre des ouvriers : 445.

A = ouvriers recevant seulement un salaire en argent .	122	
B = ouvriers recevant, outre leur salaire, la nourriture.	64	
C = ouvriers recevant, outre leur salaire, la nourriture et le logement	243	445
D = ouvriers ne recevant aucun salaire	16	

Cela étant donné, cherchons les fréquences dernières du troisième ordre ; nous avons :

$$(ABC) = 429\ (122 + 64 + 243) \qquad (\alpha BC) = 307\ (64 + 243)$$
$$(AB\gamma) = 186\ (122 + 64) \qquad (\alpha B\gamma) = 64\ (64)$$
$$(A\beta C) = 365\ (122 + 243) \qquad (\alpha\beta C) = 243\ (243)$$
$$(A\beta\gamma) = 122\ (122) \qquad (\alpha\beta\gamma) = 16\ (16)$$

Si l'on cherche la fréquence de premier ordre d'un attribut, il n'y a qu'à réunir les fréquences des classes dans lesquelles figure cet attribut. Soit à rechercher la fréquence de (A) :

$$(ABC) + AB\gamma + (A\beta C) + (A\beta\gamma) = (A) = 1102\ (429 + 186 + 365 + 122).$$

De même pour AB, ABC, AC.

$$(ABC) + (AB\gamma) = (AB) = 615\ (429 + 186)\ ;$$
$$(ABC) \qquad = (ABC) = 429\ (122 + 64 + 243)\ ;$$
$$(ABC) — (A\beta C) = (AC) = 794\ (429 + 365).$$

D'après ce qui précède, le nombre total des observations est :

$$N = 1732\ (429 + 186 + 365 + 122 + 307 + 64 + 243 + 16).$$

On a donc, en ne considérant que les classes positives :

N = 1732.	(AB) = 615.
(A) = 1102.	(AC) = 794.
(B) = 986.	(BC) = 736.
(C) = 1344.	(ABC) = 429.

Analysons la classe de dernière fréquence (A).

Elle se compose du nombre des attributs relevés dans chaque groupe de dernière fréquence où se remarque la présence de la classe (A) (ouvriers recevant un salaire en argent). En voici le détail :

(ABC)	Ouvriers payés en argent . . . Ouvriers payés en argent plus la nourriture . . . . Ouvriers payés en argent, plus la nourriture et plus le logement.	(ABC) = 429 (122 + 64 + 243)	(A) (1102)
(ABγ)	Ouvriers payés en argent . . . Ouvriers payés en argent plus la nourriture . . . . (Non compris ceux qui reçoivent de plus le logement)	(ABγ) = 186 (122 + 64)	
(AβC)	Ouvriers payés en argent . . . Ouvriers payés en argent plus la nourriture, plus le logement . (Non compris ceux qui reçoivent la nourriture et un salaire) . .	(AβC) = 365 (122 + 243)	
(Aβγ)	Ouvriers payés en argent . . . (Non compris ceux qui reçoivent la nourriture et un salaire) . . (Non compris ceux qui reçoivent un salaire, la nourriture et le logement)	(Aβγ) = 122 (122)	

143. Stanley Jevons fait remarquer (1) que le processus de développement ci-dessus est la principale particularité

(1) Cfr. Stanley (Jevons) : *On a general system, etc.* (loc. cit.), pp. 179-180.

du système de logique du professeur Boole et qu'il dépend d'une loi fondamentale de l'intelligence que Jevons appelle loi de dualité.

Soient les termes A et B; il est nécessairement vrai, d'après cette loi, que :

A est B ou n'est pas B.

Ce qui s'exprime, dans la notation proposée par Jevons, par la formule :

$$A = A\,B \cdot\!/\!\cdot A\,\beta.$$

En faisant entrer un terme nouveau, C, il est sûr que :

$$A = A\,C \cdot\!/\!\cdot A\gamma.$$

En combinant les deux développements, on arrive à la formule :

$$(A) = (A\,B\,C) \cdot\!/\!\cdot (A\,B\,\gamma) \cdot\!/\!\cdot (A\,\beta\,C) \cdot\!/\!\cdot (A\,\beta\,\gamma).$$

La classe de dernière fréquence (A), exprimée numériquement, signifie donc que, quels que soient les arrangements auxquels on s'arrête successivement entre les différents attributs exprimés par A, B, C, on n'arrivera jamais à un résultat final dans lequel les caractères additionnés des A seront supérieurs à 1102. Ce chiffre est une expression intégrale pour la classe de fréquence (A).

La notation négative exprimée par les caractères de l'alphabet grec est très commode pour marquer d'une façon simple une notion assez complexe. De plus, elle permet de résoudre facilement des formules qui, par voie de substitution, reconstituent les classes de fréquence positive.

Ainsi :

$(AB\gamma) = (AB) - (ABC) = 615 - 429 = 186.$

$(A\beta\gamma) = (A) - (AC) - (AB\gamma) = 1102 - 794 - 186 = 122.$

$(\alpha\beta\gamma) = N - (B) - (C) + (BC) - (A\beta\gamma) = 1732 - 986 - 1344 + 736 - 122 = 2468 - 2452 = 16.$

$(A\beta C) = (AC) - (ABC) = 794 - 429 = 365.$

144. Les formules données plus haut ne sont pas seulement utiles pour déterminer le nombre de classes possible qui existe dans des conditions logiques déterminées, elles peuvent aussi contribuer à faciliter le calcul des probabilités comme le montre Stanley Jevons par l'exemple suivant :

Soient p la probabilité de l'arrivée de A et q celle de l'arrivée de B ; pq est la probabilité de l'arrivée simultanée de AB ; la probabilité que A n'arrivera pas (ou celle de l'arrivée de α) est :

$$\text{pour } \alpha,\ 1 - p ; \text{ pour } \beta,\ 1 - q.$$

Donc, on a :

Probabilité de AB — pq;
Id. de $A\beta = p(1 - q)$;
Id. de $\alpha B = (1 - p)q$;
Id. de $\alpha\beta = (1 - p)(1 - q)$.

C'est encore par une application des règles logiques qui précèdent qu'on arrive à déterminer le nombre de combinaisons auxquelles peuvent donner lieu un nombre d'éléments donnés réunis dans un cadre statistique.

Le nombre d'attributs à combiner est par exemple 6.

On veut combiner ces attributs deux à deux ; on a une classe de l'ordre 2.

Ordre 2. $\frac{n(n-1)}{1 \cdot 2} = \frac{6 \cdot 5}{1 \cdot 2} = \frac{30}{2} = 15$ combinaisons possibles.

On désire combiner ces attributs trois à trois. On a donc :

Ordre 3. $\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} = \frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3} = \frac{120}{6} = 20$ combinaisons possibles.

Si l'on voulait arriver à une combinaison quatre par quatre, on aurait :

Ordre 4. $\frac{n(n-1)(n-2)(n-3)}{1.2.3.4} = \frac{6.5.4.3}{1.2.3.4} = \frac{360}{24} = 15$ combinaisons possibles.

145. Les exemples de tableaux statistiques d'une complexité croissante sont fort faciles à trouver. Ces tableaux, non seulement présentent ensemble plusieurs données, mais, de plus, les combinent en fonction les unes des autres, comme les tableaux qui groupent les causes de décès en relation avec l'âge, l'état civil, la profession, ou les statistiques de décès qui cherchent les variations du taux de mortalité en relation avec la période de l'année, la température, la pression barométrique, etc. Malgré les exemples fort nombreux qui existent, et simplement pour éviter au lecteur la fatigue de les rechercher, nous reproduisons ci-après quelques types de tableaux en les rangeant selon l'ordre de leur complexité croissante.

I. — ANGLETERRE.

Nombre des ouvriers occupés dans les industries textiles (toutes industries réunies).

ANNÉES	Nombre d'ouvriers occupés	ANNÉES	Nombre d'ouvriers occupés
1896	1,077,687	1901	1,029,353
1897	1,051,564	1904	1,026,378
1898	1,036,570	1907	1,087,223

Le nombre des ouvriers dans les industries textiles est une donnée d'un caractère général qu'on désire aussitôt développer en indiquant en regard le nombre d'ouvriers et d'ouvrières :

II. — ANGLETERRE.

Nombre des ouvriers occupés dans les industries textiles (toutes industries réunies).

ANNÉES	NOMBRE DES OUVRIERS OCCUPÉS		
	Hommes	Femmes	Total
1896	412,841	664,846	1,077,687
1897	396,851	654,713	1,051,564
1898	387,583	648,987	1,036,570
1901	379,211	650,142	1,029,353
1904	382,835	643,543	1,026,378
1907	407,360	679,863	1,087,223

Parmi ces ouvriers, on peut établir des divisions d'âges conformes aux prescriptions des lois sur la réglementation du travail. (*Voir tableau III, page suivante.*)

III. — ANGLETERRE.

Nombre des ouvriers occupés dans les industries textiles (toutes industries réunies).

ANNÉES	NOMBRE DES OUVRIERS OCCUPÉS											
	DE MOINS DE 14 ANS			DE PLUS DE 14 ET DE MOINS DE 18 ANS			DE PLUS DE 18 ANS			PERSONNEL OUVRIER TOTAL		
	Garçons	Filles	Total	Hommes	Femmes	Total	Hommes	Femmes	Total	Hommes	Femmes	Total
1896	24,302	28,954	53,256	82,383	153,862	236,245	306,156	482,030	788,186	412,841	664,846	1,077,687
1897	22,074	26,963	49,037	78,719	152,583	231,302	296,058	475,167	771,225	396,851	654,713	1,051,564
1898	20,451	24,796	45,247	76,335	151,604	227,939	290,797	472,587	763,384	387,583	648,987	1,036,570
1901	16,898	19,613	36,511	71,707	148,888	220,595	290,606	481,641	772,247	379,211	650,142	1,029,353
1904	14,568	17,176	31,744	70,965	137,038	208,003	297,302	489,329	786,631	382,835	643,543	1,026,378
1907	15,107	17,540	32,647	81,270	157,502	238,772	310,983	504,821	815,804	407,360	679,863	1,087,223

Au lieu de considérer toutes les industries réunies, la statistique peut aussi relever séparément les données qui concernent chaque branche d'industrie en particulier et présenter à part les chiffres relatifs à l'industrie du coton, de la laine (cardée et peignée), du jute, du lin et du chanvre, de la soie, de la bonneterie, de la dentelle, des tissus élastiques, crin, etc.

Le tableau définitif prend alors la forme ci-après. (*Voir tableau IV, page suivante.*)

---

(1) D'après *Statistical Abstract for the United Kingdom*, London, 1913, no 81, pp. 352-53.

IV. — ANGLETERRE.

Nombre des ouvriers occupés dans les industries textiles (répartition par âge, par sexe et par industrie).

MATIÈRE PREMIÈRE EMPLOYÉE	TRAVAILLANT A DEMI-TEMPS — Enfants de moins de 14 ans			TRAVAILLANT A PLEINES JOURNÉES						PERSONNEL OUVRIER TOTAL			MATIÈRE PREMIÈRE EMPLOYÉE
				En dessous de 18 ans			De plus de 18 ans						
	Masc.	Féminin	Total	Masc.	Féminin	Total	Masc.	Féminin	Total	Masc.	Féminin	Total	
COTON.													COTON.
1896	13,185	16,321	29,506	41,355	80,061	121,416	149.146	232,852	381,998	203,686	329,234	532,920	**1896**
1897	12,297	15,246	27,543	40,034	79,580	119,614	147,245	232,821	380,066	199,576	327,647	527,223	**1897**
1898	11,697	14,334	26,031	39,362	80,187	119,549	146,642	233,885	380,527	197,701	328,406	526,107	**1898**
1901	9,780	11,173	20,953	37,002	77,631	114,633	147,048	239,989	387,037	193,830	328,793	522,623	**1901**
1904	8,131	9,520	17,651	37,338	71,975	109,313	150,952	245,114	396,066	196,421	326,609	523,630	**1904**
1907	8,862	10,189	19,051	45,766	85,637	131,403	163,114	263,252	426,366	217,742	359,078	576,820	**1907**
LAINE (peignée et cardée) **etc.**													**LAINE** (peignée et cardée) **etc.**

Une cinquième forme de présentation peut être introduite en distinguant les comtés ou les localités. Elle est très usuelle car la présentation sur la base géographique est des plus féconde en aperçus.

146. Etant donnés ces éléments, le statisticien a le choix entre différentes combinaisons pour présenter les résultats de la statistique. A son dernier stade, le relevé ci-dessus se compose de quatre éléments : la date du relevé, la nature de l'industrie, l'âge des ouvriers et leur sexe. Si ces éléments sont associés deux à deux, on aura :

$$\frac{n(n-1)}{1 \cdot 2} = \frac{4 \cdot 3}{1 \cdot 2} = \frac{12}{2} = 6 \text{ combinaisons.}$$

S'ils sont associés trois à trois, on aura :

$$\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} = \frac{4 \cdot 3 \cdot 2}{1 \cdot 2 \cdot 3} = \frac{24}{6} = 4 \text{ combinaisons.}$$

Dans le Census anglais, par exemple, on trouve 12 attributs différents, ce qui fait, dit M. Bowley (1), qu'en les combinant ensemble, on a le choix entre 66 façons de tabuler différentes quand on prend 2 éléments ensemble, 220 quand on en prend 3, 495 quand on en prend 4. En effet :

$$\frac{n(n-1)}{1 \cdot 2} = \frac{12 \cdot 11}{1 \cdot 2} = \frac{132}{2} = 66$$

$$\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} = \frac{12 \cdot 11 \cdot 10}{1 \cdot 2 \cdot 3} = \frac{1320}{6} = 220$$

$$\frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4} = \frac{12 \cdot 11 \cdot 10 \cdot 9}{1 \cdot 2 \cdot 3 \cdot 4} = \frac{11880}{24} = 495$$

Ainsi, représentant la nature de l'industrie par I, l'âge des ouvriers par A, leur sexe par S, la date du relevé par D, les résultats du tableau (IV) peuvent être présentés comme ceci :

			A_1			
			D_1	D_2	D_3	D_4
I_1	S	S_1				
		S_2				

(1) Bowley (A.), *Elements of statistics*, London, 1902, p. 78.

ou encore de cette manière :

I_1

D_1 D_2 D_3 D_4

A_1, A_2, A_3 } S { S_1, S_2

On peut encore procéder d'une autre façon :

S_1: I_1 I_2 I_3 I_4 — S_2: I_1 I_2 I_3 I_4

A_1 { D_1, D_2, D_3

Il est aussi possible de rapprocher les S en envisageant séparément les I :

I_1: S_1 S_2 — I_2: S_1 S_2

A_1 { D_1, D_2, D_3

Etc...

Entre ces combinaisons et toutes les autres possibles, on choisira évidemment celle qui paraîtra le mieux à même de répondre de la façon la plus adéquate à la question qu'on veut élucider. La forme générale la plus simple est celle qui consiste à énumérer dans la colonne de gauche les industries et les dates du relevé, comme dans le tableau IV ci-dessus. L'indication de la localité prend souvent la place des dates, car il est rare qu'on compare des situations à des dates différentes dans la même localité.

III. — Des classifications statistiques.

147. Les classifications sont, en statistique, d'une importance exceptionnelle, parce que d'elles dépend, en bonne partie, le degré de pénétration et d'utilisation des résultats du relevé. Il ne servirait de rien d'avoir, au cours du relevé, respecté les règles fondamentales de l'observation selon le temps et l'espace, si ces distinctions essentielles ne se retrouvaient pas dans les classifications suivant lesquelles les unités sont ordonnées. Il existe donc un lien intime entre les règles qui président au classement et celles qui concernent le relevé : c'est un point qui a semblé échapper à l'attention de plusieurs statisticiens et qu'il importe cependant de poser dès le début.

Classer, c'est répartir les unités du phénomène en groupes ou sous-groupes d'après leurs caractères, et c'est, d'autre part, observer les différences qu'elles présentent quand elles sont groupées selon un caractère donné.

Soit une statistique des décès classés d'après leurs causes : on a d'abord à établir une classification des causes de décès, comme celle basée sur le siège des maladies, par exemple. Ensuite, en examinant le nombre de décès causés par chaque maladie, on devra répartir ce nombre selon certaines divisions, telles que l'âge et l'état civil des décédés.

Il suit de là que les classements en statistique sont de deux espèces : 1° ceux qui visent les différents caractères que le phénomène peut revêtir quand on le considère sous un aspect spécial; 2° ceux qui partagent en catégories diverses les unités observées sous un point de vue particulier et faisant partie d'un ensemble : c'est la distribution par fréquence.

Nous exposerons successivement les règles générales applicables aux classifications comme à la distribution par

fréquence; ensuite, nous verrons quelles sont les applications logiques qui découlent de l'étude des tableaux à double entrée dans lesquels se combinent les deux principes de classement indiqués ci-dessus, tableaux auxquels le professeur Pearson a donné le nom de tableaux de contingence (contingency-tables).

148. Les classifications proprement dites doivent répondre à une idée générale, la plus simple et la plus étendue qu'il se peut, de façon à comprendre toutes les manifestations du phénomène étudié. Cette idée générale est susceptible de revêtir un grand nombre de formes; il est rare qu'un relevé statistique ne comporte qu'une seule classification; d'ordinaire, il en faut deux ou trois pour mettre en évidence les divers aspects sous lesquels le phénomène demande à être étudié. Les statistiques démographiques utilisent des classifications basées sur des critères appropriés aux faits qu'elles considèrent : la répartition géographique, à l'aide de laquelle peut se déterminer l'influence du milieu; les répartitions basées sur l'âge des recensés dans les dénombrements de la population, sur l'âge des époux au moment du mariage ou du divorce, sur l'âge des habitants classés d'après leur degré d'instruction; le partage des décès s'opère naturellement d'après l'âge, comme aussi selon les maladies ou l'état civil ou la profession des décédés; dans les statistiques économiques on adoptera comme base des classifications usuelles l'industrie ou le métier; dans les statistiques commerciales, la nature des produits échangés, les provenances et les destinations des échanges commerciaux. Ainsi, chaque espèce de statistique a ses principes de classification particuliers dont elle ne s'éloigne guère. Il faut et il suffit que ce principe soit assez large pour couvrir toutes les manifestations du phénomène, de même que le relevé doit s'étendre à toutes ces manifestations, en quelque lieu qu'elles se produisent. Le principe de classification doit être approprié au but particulier de

chaque statistique; c'est le talent du statisticien de reconnaître entre des principes différents et également bons au premier aspect, celui qui pourra le mieux faire ressortir les caractères essentiels du phénomène, en même temps qu'il s'adaptera le mieux aux limites de fréquence établies d'autre part.

149. Dans les classifications statistiques, deux difficultés opposées sont à vaincre. Les nomenclatures trop brèves embrassent dans une même classe ou rubrique des choses extrêmement différentes, n'ayant souvent entre elles d'autres rapports que la désignation générale sous laquelle il est convenu de les réunir. Ainsi, la statistique commerciale de la Hollande, qui ne comprend qu'un nombre assez limité de rubriques, se termine par une classe comprenant toutes les marchandises autres que celles énumérées précédemment; dans une classe de la statistique commerciale belge (merceries-quincailleries) ne sont pas réunies moins de cent quarante-six espèces de marchandises des plus disparates. Il convient d'éviter dans les classifications les rubriques intitulées : « divers ». L'obligation où l'on se trouve d'y recourir prouve ou que l'énumération est incomplète, ou que le principe de la classification est mal choisi en ce qu'il ne peut s'adapter à tous les cas qui se présentent.

D'un autre côté, les nomenclatures trop étendues ne sont pas sans inconvénients : encombrantes et malaisées à consulter, elles présentent de sérieuses difficultés au moment de l'interprétation en ce qu'elles ne groupent plus, sous chacune de leurs rubriques, que des faits trop peu nombreux. Cependant, réserves faites quant à l'abus de la spécialisation, les classifications doivent, en principe, être aussi détaillées que possible, afin de permettre de conserver aux faits observés leur caractère individuel constaté lors du relevé.

150. Il est souvent dangereux de recourir, dans les classifications statistiques, à des principes empruntés au clas-

sement des sciences. Les vues théoriques qui dominent à un moment donné sont sujettes à se modifier, de sorte qu'après un certain temps, on est fort étonné de trouver réunies des rubriques de classement dont l'accouplement ferait sourire si on ne se rappelait qu'il fut fondé autrefois sur des vues généralement acceptées. L'inconvénient qui résulte de ce fait est qu'il n'est guère possible de conserver de telles bases quand elles sont surannées et que, dès lors, le grand avantage qui résulte des comparaisons se trouve annihilé.

Une classification nosologique des causes de décès, qui eût été élaborée il y a quelque cinquante ans, nous étonnerait, aujourd'hui, par les rapprochements bizarres qu'elle renfermerait ; les théories médicales se sont modifiées et les opinions sur les causes des maladies ne sont plus ce qu'elles étaient autrefois. William Farr avait donc raison quand il déclarait préférer à la classification nosologique celle, beaucoup plus terre à terre, basée sur le siège de la maladie. Au moins celle-ci a-t-elle des chances de durée (1).

On a voulu aussi baser les classifications de l'industrie d'après la matière première que les industries mettent en œuvre et l'on a repris, dans ce but, l'antique classification des trois règnes : minéral, végétal et animal. Ce principe de classification conduit aux rapprochements les plus bizarres : ainsi, l'on voit figurer ensemble dans une même division, la fabrication du gaz d'éclairage, le travail et la préparation des pierres, l'industrie du lapidaire, le raffinage du sel et la fabrication du verre à vitre ; cette association se justifie par le fait que toutes ces industries emploient des matières minérales. Le principe de classification est d'ailleurs insuffisant, car il ne peut servir pour les industries qui utilisent à la fois des matières premières empruntées à deux ou aux trois règnes. Où faut-il classer la construction des wagons ? et la construction des bâtiments ?

---

(1) BERTILLON, *Traité élémentaire de statistique administrative*, Paris, 1886.

Il est bien préférable de s'en tenir à des classifications plus simples et plus vraies, basées sur l'usage. Tout le monde sait ce que l'on veut dire lorsqu'on parle de l'industrie des métaux, de la construction, du vêtement; cet avantage n'est pas mince. Il ne suffit pas que le principe de classification soit — s'il est possible — irréprochable, il faut encore qu'il soit compris. Admettons, comme Stanley Jevons le suppose (1), que nous puissions classer tous les corps en raison de leur constitution atomique. Sans conteste, une pareille classification serait essentiellement scientifique et peut-être serait-elle capable de suggérer des rapprochements fort intéressants. Le malheur est, qu'en dehors d'un petit nombre de spécialistes, elle ne pourrait servir à personne.

151. Les classifications dont nous venons de parler sont toutes basées sur un ordre idéologique quelconque. Pour obvier aux difficultés signalées, certains techniciens ont parfois essayé de nomenclatures basées sur l'ordre alphabétique. Le remède est pire que le mal. Les classifications alphabétiques sont détestables, car le chercheur ne sait jamais si telle dénomination comprend toutes les variétés qui peuvent exister et, pour atteindre des résultats sûrs, le lecteur est forcé de se livrer à des recherches aussi longues que fastidieuses; aucune totalisation systématique ne peut se faire si l'on s'en tient rigoureusement à l'ordre alphabétique; enfin, les comparaisons internationales — qu'il ne faut jamais perdre de vue, tout en se montrant fort prudent dans la pratique — sont décidément impossibles avec ce système. M. J. Bertillon a dit avec esprit que l'ordre alphabétique n'était qu'une forme du désordre : il a parfaitement raison.

152. Indépendamment des considérations qui précèdent, il convient d'introduire dans la matière qui nous occupe,

(1) Stanley (Jevons), *Principles of Science. Classification.*

une dernière distinction : celle entre les classifications homogènes et les hétérogènes. On donne le nom d'homogènes aux classifications qui se rapportent à un seul phénomène considéré dans ses diverses parties. Ainsi, une classification des âges par périodes de cinq années est homogène, parce qu'on ne considère que le phénomène âge. Mais, dans un classement des habitants d'un pays d'après leur profession, des données hétérogènes doivent nécessairement être introduites. Prenons comme exemple la nomenclature méthodique des industries et professions adoptée pour le recensement français. Nous avons : hors section, les personnes ne vivant pas de l'exercice d'une profession proprement dite ou exerçant une profession mal déterminée ; viennent ensuite les sections : 1° pêche ; 2° forêts et agriculture ; 3° industries extractives ; 4° industries de transformation ; 5° manutention et transport ; 6° commerce, banque ; 7° professions libérales ; 8° soins personnels, service domestique ; 9° services de l'Etat, des départements ou des communes. Cette classification est évidemment hétérogène et elle ne peut avoir un autre caractère.

Il reste une dernière remarque à présenter : elle a trait aux comparaisons statistiques et au rôle que jouent, dans ces comparaisons, les classifications. Il est rare qu'après avoir arrêté une classification, on ne doive pas, quelques années après, y introduire des changements. La vie économique, entre autres, ne se conçoit pas sans changement, de telle sorte qu'un classement des industries est en perpétuelle instance de revision. A un moindre degré, il en est de même des classifications de maladies, auxquelles la science médicale travaille constamment à apporter des modifications. S'il s'agit donc de comparer, à des dates différentes, des faits classés d'après un certain ordre, trop de soins ne pourront être apportés à la vérification des rubriques de classement de façon à ne comparer que ce qui est rigoureusement comparable ; les faits nouveaux seront mis à part, comme aussi les faits qui ont cessé d'être constatés.

Cette précaution ne suffit pas encore : il convient de s'assurer si les mêmes désignations s'appliquent encore aux mêmes faits. Sans que les dénominations des rubriques aient été changées, leur contenu peut avoir varié dans une sensible mesure. L'auteur de la comparaison ne doit pas s'y laisser tromper et il a le devoir de signaler au lecteur les changements internes dont il a connaissance.

153. Les classifications disposent les faits d'après leur nature, en tenant compte de leurs ressemblances et de leurs différences. A chaque catégorie correspond un certain nombre de manifestations du phénomène ; il faut non seulement observer leur nombre, mais les répartir d'après leur importance, leur intensité, les circonstances dans lesquelles elles se produisent. Ceci est la mission propre du tableau statistique.

Dans le dépouillement, on compte unité par unité, mais pour présenter les résultats du comptage il est nécessaire, le plus souvent, de les réunir par groupes plus ou moins étendus. Dans une statistique des âges, le statisticien n'énumérera pas tous les âges, par mois, ni même par années, mais composera des groupes d'âges (moins de 1 an — de 1 à 5 ans — de 5 à 10 ans, etc.) où viendront se réunir tous les résultats partiels compris dans les limites de chaque groupe. Autre exemple : il s'agit d'opérer la répartition des entreprises trouvées lors d'un recensement industriel, d'après le nombre des ouvriers que chacune d'elles emploie. Toutes les entreprises qui n'ont pas exactement le même nombre d'ouvriers ne peuvent être mises à part, sinon la classification comporterait plusieurs centaines de termes (établissements de 1, 2, 3... 297, 298, 299, 300, 301... ouvriers). Aussi présente-t-on les résultats du dépouillement suivant certaines catégories : moins de 5 ouvriers, de 5 à 10, de 11 à 20, de 21 à 50, de 51 à 100, de 101 à 200, etc.

Le groupement est requis par des nécessités scientifiques et pratiques. Scientifiques : on éprouverait une extrême

difficulté à dégager quelque résultat que ce soit d'une statistique comportant une classification aussi étendue et il faudrait en venir à grouper les données pratiques : les statistiques atteindraient des dimensions telles qu'il faudrait renoncer à les publier : beaucoup de colonnes resteraient vides; avec des divisions multipliées, les chances d'erreur dans le dépouillement seraient augmentées.

154. Jusqu'à quel point convient-il de réunir, de grouper les unités comptées par le dépouillement? Il n'y a pas, à cet égard, de règle absolue; les circonstances, l'étude approfondie de la matière doivent, en cette circonstance, inspirer les solutions à choisir. Cependant, quelques recommandations générales peuvent être formulées. L'auteur du travail prendra égard, par exemple, au sujet de la recherche statistique : s'il s'agit de grouper une population par âge, des divisions quinquennales seront, somme toute, suffisantes, sauf pour les âges extrêmes où il y a intérêt à isoler des cas qui méritent une étude spéciale. Lorsque la classification par âges doit servir à une statistique des décès, les divisions, au début, seront assez courtes pour pouvoir isoler la mortalité infantile. Si la statistique mortuaire est destinée à faire ressortir l'influence des saisons et des facteurs qui y ont rapport, les divisions quinquennales d'âges seront unies à des divisions détaillées de l'année (mois par mois).

Ce premier point ayant été déterminé avec soin, il faut arrêter le nombre de divisions à adopter. Il importe de distinguer entre les variables qui ne procèdent que par degrés de faible étendue, de celles où les écarts sont plus considérables. Les fleurs de beaucoup d'espèces botaniques n'ont pas toujours le même nombre de pétales; pour observer une variable de ce genre, il va de soi que la division à adopter sera l'unité. Lorsque les variations comportent un plus grand nombre de degrés, du point le plus bas au point le plus élevé, on pourra adopter des divisions plus larges

que dans le cas précédent. Toutefois, il convient d'éviter toujours de trop condenser la matière; s'il y a moins de dix divisions, il sera difficile d'éviter que la statistique présente une fâcheuse imprécision, car chaque classe renfermera alors des unités sensiblement différentes. D'autre part, on ne peut multiplier trop les divisions, car alors on aboutirait à un tableau illisible; il est fort désirable, comme nous le verrons plus loin, que les tableaux statistiques puissent tenir sur une double page du volume.

155. Le point de départ des divisions est aussi une question qui mérite réflexion. Vaut-il mieux adopter des divisions normales, commençant par l'unité simple et finissant avec elle, comme de 12 à 13 — 13 à 14, etc., ou bien est-il préférable de choisir l'unité augmentée d'une fraction, 1/2 par exemple, comme 12.5 à 13.5 — 13.5 à 14.5? Toutes les unités ne se prêtent pas à une division de ce genre, ou elles ne s'y prêtent pas dans tous les cas. Cette réserve faite, il y a souvent avantage à adopter des divisions telles que : 12.5 à 13.5 — 13.5 à 14.5, afin d'éviter de rendre trop sensibles les erreurs qui se marquent fréquemment aux degrés de l'échelle correspondant à des nombres ronds (20, 30, 40); les exagérations à ces degrés sont fréquentes dans les recensements de la population et existent aussi dans d'autres dénombrements.

Il est important d'adopter, dans les classifications, des divisions qui ne prêtent à aucune équivoque. Dans un relevé où il s'agit, par exemple, de sommes d'argent classées d'après leur importance, le classement se fait souvent comme ceci : moins de 1,000 francs; de 1,000 à 2,000 francs; de 2,000 à 3,000 francs, etc. Il est plus exact de dire : de 1,000 à 1,999 francs, de 2,000 à 2,999 francs, etc. La limite à laquelle on arrête les divisions indique le degré de précision du relevé. Dans l'exemple donné ci-dessus, cette limite est le franc; si elle était le centime, l'énumération devrait

être faite ainsi : de 1,000 à 1,999.99 francs; de 2,000 à 2,999.99 francs, etc.

Pour se conformer à la théorie, les divisions numériques adoptées dans les diverses classifications devraient être égales : si l'on admet au début que les échelons seront distants de dix degrés, il est évident que, théoriquement, cette convention devrait être respectée jusqu'à la fin. C'est là une exigence que les statistiques officielles ne sont pas à même, le plus souvent, de respecter et cela pour un grand nombre de raisons : d'abord, parce que ce plan aurait pour effet d'augmenter beaucoup l'étendue des publications et, par conséquent, leur coût; ensuite, parce que le dépouillement serait plus long et plus coûteux; enfin, parce que beaucoup de colonnes resteraient vides ou ne contiendraient que de rares unités. Très souvent, les divisions des échelles sont donc inégales. A certains égards, il en résulte de sérieux inconvénients; beaucoup de calculs de la statistique mathématique supposent qu'on puisse tabler sur des divisions égales. On devrait s'attendre par exemple à observer une diminution régulière dans les chiffres d'un relevé et l'on constate au contraire des ressauts qui rompent l'uniformité de la descente : ils sont dus le plus souvent à l'augmentation subite du nombre de degrés compris dans une nouvelle classe. Un moyen simple de remédier à cette irrégularité consiste à diviser les résultats de chaque classe par le nombre de degrés qu'elle renferme, lorsque ce nombre est supérieur à la grandeur admise comme point de départ.

IV. — Préparation des tableaux.

156. Quand un phénomène est décrit à l'aide de données statistiques exposées selon un double système de classification, nous avons ce qu'on appelle communément un tableau à double entrée. Ce genre de tableau statistique est des plus utile en ce que les données n'y sont pas présentées suivant

un plan unique (auquel cas on a simplement une description du phénomène) mais qu'elles sont mises en fonction les unes des autres, ce qui favorise beaucoup la recherche des relations causales et projette une vive clarté sur l'essence du phénomène lui-même. Dans les tableaux à entrée simple, on énumère, on expose, on totalise; le tableau à double entrée participe déjà de l'interprétation, tant il met en relief les caractéristiques du phénomène lié à un certain ordre de causes. Une répartition de la population d'après les professions a un aspect tout différent selon qu'elle suit simplement la classification des industries ou qu'elle combine la notion « profession » avec la notion « âge » ou « origine ». Entre les divisions des tableaux à double entrée se remarquent des relations logiques de l'ordre de celles que nous avons exposées plus haut, en ce qui concerne les classes de dernière fréquence, comme le montre l'exemple suivant :

Nombre de bâtiments existant en Belgique (1910) :

IMPORTANCE DES COMMUNES	NOMBRE DE BATIMENTS AU 31 DECEMBRE 1910			
	Maisons proprement dites destinées à l'habitation (habitées ou non)	Bâtiments de toute nature non destinés à l'habitation mais où demeurent une ou plusieurs personnes	Autres bâtiments non destinés à l'habitation et non habités	TOTAL
(1)	(2)	(3)	(4)	(5)
U. — Communes de 20,000 habitants et plus. .	366,894	4,274	10,140	381,308
M¹. — Communes de 5,000 à moins de 20,000 h.	440,682	2,779	15,037	458,498
M². — Communes de 2,000 à moins de 5,000 h.	343,627	2,347	9,538	355,512
R. — Communes de moins de 2,000 habitants .	385,133	3,533	14,801	403,467
TOTAUX. . .	1,536,336	12,933	49,516	1,598,785

Pour la facilité de l'analyse, désignons par A la colonne 2, par B la colonne 3, par C la colonne 4, par D la colonne 5; appelons U les communes urbaines de 20,000 habitants et plus, R les communes rurales de moins de 2,000 habitants, M^1 les communes de 5,000 à moins de 20,000 habitants, M^2 celles de 2,000 à moins de 5,000 habitants... Nous avons :

$$\text{pour U : } \frac{A}{D} = 96.22 \quad \frac{C}{A} = 2.76 \quad \frac{C}{AB} = 2.73 \quad \frac{B}{A} = 1.16$$

$$\text{pour R : } \frac{A}{D} = 95.46 \quad \frac{C}{A} = 3.84 \quad \frac{C}{AB} = 3.81 \quad \frac{B}{A} = 0.92$$

$$\text{pour } M^1 : \frac{A}{D} = 96.11 \quad \frac{C}{A} = 3.41 \quad \frac{C}{AB} = 3.39 \quad \frac{B}{A} = 0.63$$

$$\text{pour } M^2 : \frac{A}{D} = 96\ 66 \quad \frac{C}{A} = 2.78 \quad \frac{C}{AB} = 2.76 \quad \frac{B}{A} = 0.58$$

De quoi l'on peut conclure : 1° que la proportion des maisons destinées à l'habitation est pour ainsi dire la même dans les localités urbaines et les communes importantes sans que le facteur population joue à cet égard un rôle sensible; il n'y a que dans les communes rurales que la proportion des bâtiments destinés à l'habitation soit plus faible;

2° Les bâtiments non destinés à l'habitation et non habités sont plus nombreux dans les campagnes que partout ailleurs; la proportion diminue dans les localités secondaires pour arriver au minimum dans les villes importantes, mais cette diminution n'est pas parallèle à la marche du chiffre de la population, du moins dans le classement proposé;

3° C'est dans les grandes villes que la proportion des bâtiments non destinés à l'habitation mais où demeurent une ou plusieurs personnes est la plus forte; elle est moindre dans les campagnes, mais elle est surtout très faible dans les petites villes et les grosses communes;

4° Le rapport des bâtiments non destinés à l'habitation à ceux qui, à un titre quelconque, servent d'habitation est le

plus élevé dans les communes rurales, puis dans les villes de second rang (5.000 à 20.000) ; il est le plus faible dans les villes et un peu plus élevé dans les localités de 2,000 à moins de 5,000 habitants.

CHAPITRE II

Exécution du dépouillement

I. — Organisation du dépouillement.

157. Dans toute opération technique, la valeur des résultats est en raison directe de la perfection des moyens employés. Il y a donc un intérêt primordial à ne recourir qu'à des méthodes éprouvées tout en étant décidé à accueillir sans hésitation les améliorations dues aux progrès des sciences ; il faut savoir être assez conservateur pour ne pas sacrifier à la légère une organisation dont les résultats furent longtemps jugés satisfaisants, mais, d'autre part, rien ne serait plus désastreux que l'entêtement à ne pas s'écarter des voies suivies jusque-là et l'obstination à ne pas reconnaître les progrès accomplis. C'est de cette double pensée qu'il convient de s'inspirer en arrêtant les méthodes à employer dans l'exécution du dépouillement.

Le dépouillement peut être exécuté de deux façons : par la méthode décentralisée et par la méthode centralisée.

La méthode décentralisée est celle qui consiste à confier à des autorités administratives inférieures le soin, non seulement de recueillir les renseignements statistiques, mais de les réunir, de les classer et de les dépouiller suivant le plan et d'après les instructions transmis par l'organe central. Dans ce système, le travail de l'organisme central consiste principalement dans des contrôles formels plutôt qu'essentiels et à mettre bout à bout les divers relevés partiels que lui transmettent les autorités subordonnées. Dans la méthode centralisée, au contraire, aucun dépouillement n'est effectué qu'au centre ; c'est à peine si les agents recen-

seurs et les administrations communales sont chargés de quelques classements préalables n'ayant rien de commun avec le dépouillement proprement dit.

Lorsque la méthode décentralisée est appliquée dans toute sa rigueur, le dépouillement ne se fait pas en une étape, mais en plusieurs, aussi nombreuses que les degrés de l'échelle administrative qui relie l'organe exécutant à l'organe central. Tel était le système suivi en France pour le recensement général de la population avant 1901. Les mairies étaient chargées du dépouillement des données se rapportant à la commune ; les sous-préfectures réunissaient les données des communes de leur ressort, et les préfectures, à leur tour, groupaient les données par département. L'organe central recevait uniquement ces « états », objets de manipulations successives auxquelles il était resté totalement étranger ; ne possédant pas les documents originaires, il n'était pas à même de procéder à une revision quelconque du matériel statistique et du moment que les opérations arithmétiques étaient correctes, il devait s'estimer satisfait. Déjà en 1880, une commission nommée auprès du ministère de l'Intérieur, puis le Conseil supérieur de statistique, attirèrent l'attention sur les inconvénients de ce système ; en 1894, une commission instituée par le ministre du Commerce se prononça aussi en faveur du dépouillement central des bulletins du recensement des industries et professions dont elle avait à tracer le plan. Cette recommandation fut entendue en 1896 pour la partie professionnelle des bulletins de recensement. Depuis 1901, le dépouillement du recensement de la population est, en France, complètement centralisé.

158. A quel système convient-il d'accorder la préférence? La question, qui a pu paraître douteuse il y a longtemps, ne l'est plus aujourd'hui. La science et la pratique se sont prononcées en faveur du dépouillement centralisé et à l'appui de leur décision l'on apporte les arguments suivants :

1° Le dépouillement centralisé est le seul système qui permette d'introduire dans l'administration de la statistique le principe général d'organisation qui, depuis longtemps, est en vigueur dans l'industrie et dans les sciences : la spécialisation des fonctions. Celui-ci entraîne les conséquences suivantes : les aptitudes diverses des agents sont employées au mieux ; l'entraînement résultant de la répétition de la même opération augmente le rendement ; le chef de l'organisme central, surveillant tous les résultats des diverses opérations, peut corriger progressivement les défauts que le fonctionnement accuse ; un outillage perfectionné peut être mis en usage avec son maximum de rendement, etc. ;

2° Au point de vue administratif, la centralisation présente l'avantage de mettre fin aux revendications justifiées que faisaient entendre les administrations inférieures, déjà encombrées de travail, et auxquelles l'exécution des statistiques apportait un notable surcroît de besogne ;

3° Le travail statistique dans les administrations inférieures étant intermittent, il n'y a pas possibilité d'y former des employés vraiment aptes à cette tâche. Les statistiques sont dépouillées par des agents non qualifiés spécialement, n'ayant pas d'intérêt à l'exécution correcte du travail ; au contraire, le dépouillement centralisé fait passer le travail statistique aux mains d'un personnel spécialement recruté, responsable, intéressé à accomplir sa tâche le mieux possible ;

4° Il est impossible d'assurer l'uniformité des décisions concernant les classements et les autres questions, si nombreuses, qui se posent au cours du dépouillement, ailleurs que dans le système centralisé. Une telle entreprise rentre-t-elle dans cette rubrique de classement, ou cette autre ? S'agit-il d'une entreprise appartenant à l'industrie à domicile, ou à l'industrie en atelier ? De pareilles questions, qui exigent des solutions d'espèce, ne peuvent être résolues par des instructions générales si détaillées qu'on veuille les sup-

poser. Pour être homogènes, les réponses doivent être formulées par un organisme central. La solution consistant à recommander aux autorités communales et provinciales de soumettre à l'organe central les questions douteuses, est illusoire : cet expédient ferait perdre trop de temps et il n'est utilisé par personne. De même, est insuffisante à assurer l'homogénéité des décisions, la communication par l'organe central, à toutes les autorités subordonnées, des solutions qu'il a données à des difficultés qui lui ont été soumises : rien ne dit qu'on tient compte de pareilles recommandations, ni qu'elles n'arrivent pas trop tard.

5° L'utilisation complète du matériel statistique exige la confection de nombreux tableaux dérivés qu'un personnel spécial est seul à même de concevoir et d'exécuter. Ce personnel ne peut se trouver que dans un bureau central où il a acquis, par la pratique, les qualités nécessaires à sa tâche. Pour bien utiliser une statistique, il faut avoir assisté à toutes les phases du dépouillement, sinon on court le risque de commettre de fortes erreurs dans l'interprétation. Au nombre des conditions requises pour l'utilisation rationnelle des données, on peut citer l'emploi d'un outillage mécanique complet, qui ne peut se trouver que dans les bureaux d'un organisme central.

159. A ces raisons, on oppose quelques objections, mais sans une valeur décisive : 1° le bureau central sera submergé sous la masse des documents à utiliser. Réponse : il est vrai que la tâche du bureau central est beaucoup plus rude que par le passé, mais l'expérience acquise depuis plus de trente ans montre que la besogne dont il est question ne dépasse pas les forces administratives dont dispose l'Etat moderne;

2° Ecrasé par sa besogne matérielle, l'organe central ne pourra donner une attention suffisante à la correction des documents. Réponse : les faits, qui sont plus éloquents que tout le reste, prouvent exactement le contraire. Une revi-

sion sérieuse n'est possible qu'à la condition que le bureau central possède les bulletins mêmes de la statistique et qu'il soit chargé de les dépouiller. Toute l'évolution de la statistique administrative se fait en ce sens depuis de nombreuses années;

3° La correction des bulletins est malaisée à cause de l'éloignement du bureau central : il ignore les circonstances locales et ne peut s'en inspirer. Réponse : l'objection ne serait fondée que dans l'hypothèse seulement où le bureau central prendrait la charge de toutes les corrections, mais dans la plupart des cas il se borne à renvoyer aux autorités locales les bulletins défectueux (1). De nos jours, les moyens actuels de communications permettent à l'organe central de se renseigner rapidement et à peu de frais auprès des intéressés eux-mêmes : ainsi, lors du recensement belge de 1910, nous avons fait usage du téléphone d'une manière intensive pour réclamer aux chefs d'entreprise les renseignements complémentaires qu'il fallait réunir.

Nous croyons inutile de répondre à une dernière objection d'après laquelle, dans le système centralisé, le chef de l'administration manquerait de temps pour cultiver la science.

Il n'est pas douteux que le dépouillement centralisé soit une forme supérieure de l'organisation de la statistique. Aussi presque tous les pays l'ont-ils successivement adopté. Il ne reste guère de vestiges de l'organisation ancienne que dans la mission confiée, parfois encore, aux autorités locales, de classer les bulletins statistiques dans l'ordre indiqué par le bureau central. Il n'y a rien à redire à cette modalité de la décentralisation; elle a pour effet de débarrasser

(1) L'Office du Travail de Belgique, lors du recensement de l'industrie et du commerce (1910), a renvoyé, pour correction, aux autorités locales, 383,000 bulletins de recensement, sans compter les très nombreuses corrections qu'il a effectuées lui-même, d'après les renseignements réclamés directement aux chefs d'entreprise.

l'organe scientifique d'une besogne aussi longue que fastidieuse.

Les forces administratives mises à la disposition de l'organisme central pour venir à bout de la tâche énorme qui lui est assignée, sont parfois considérables. Aux Etats-Unis, lors du recensement fédéral de 1910, 650 personnes étaient occupées, au début de l'année 1910, aux travaux préparatoires; 3,800 personnes étaient au travail au mois de septembre 1910, leur nombre était réduit à 2,868 en juin 1911 et à 2.458 en septembre de la même année. Le personnel est toujours engagé à proportion du travail actuel; aussi la durée des services est parfois très limitée : 60 jours seulement dans certains cas. Le nombre des employés d'administration est très faible par rapport à celui des agents temporaires. Ainsi, il n'y avait que 24 personnes du service administratif pour 2,540 employés, 169 aides et 42 ouvriers mécaniciens occupés à la fabrique de machines à calculer annexée aux bureaux du Census (1).

II. — Méthodes de dépouillement.

160. Nous ne parlerons pas ici des procédés mécaniques, dont la description se trouve au chapitre suivant; nous envisageons uniquement les méthodes de dépouillement sous leur aspect le plus général : la méthode de pointage et la méthode de comptage par bulletins individuels ou par fiches. La méthode de pointage consiste à dépouiller chaque bulletin séparément et à marquer l'enregistrement d'une unité de chaque espèce en portant, dans la colonne réservée à ces unités, une barre verticale | et en indiquant la cinquième unité dépouillée au moyen d'une ligne oblique ~~||||~~ . Cette méthode est simplifiée dans la pratique courante : au lieu de remplir les colonnes de la feuille de dépouillement

(1) *Reports of the Department of Commercial Labor*. Washington, 1912, p. 49.

de barres verticales et obliques, on a coutume d'effectuer ce comptage sur des feuilles sommaires; on reporte, après un certain temps, ces résultats sur la feuille de dépouillement mais en les marquant en chiffres. Cette méthode présente l'avantage de conserver aux feuilles de dépouillement un aspect plus propre, de sorte que, après avoir totalisé les différents chiffres portés dans chaque colonne et les avoir inscrits à l'encre, le manuscrit peut, sans être recopié, être envoyé à l'impression.

M. Bertillon a proposé un système un peu différent du pointage. Il consiste à inscrire, au lieu de barres, autant de chiffres correspondant au numéro d'ordre de la colonne où on les inscrit. « Il est clair, dit M. Bertillon, qu'il n'est pas plus long d'écrire un 2 que de tracer un bâtonnet, et, d'autre part, cela ôtera toute chance d'erreur, car on n'aura jamais la tentation d'écrire le chiffre 2 dans la colonne des 3. » Et il ajoute : « Cette manière de faire n'est pas toujours pratique. » Nous sommes pleinement d'accord sur ce point : il est évident que ce procédé de notation perd de sa valeur à mesure qu'on s'avance dans l'ordre des nombres. Dès la dizaine, le procédé n'est plus pratique; or, nous avons vu (cfr. n° 117) qu'une dizaine de subdivisions dans l'échelle de classification des unités peut être regardée comme insuffisante, en général.

161. La méthode de dépouillement par pointage est impraticable dans le cas où il s'agit d'un matériel statistique contenu dans des registres. Elle ne s'applique qu'au type de bulletin collectif et elle suppose, pour son application rationnelle, qu'il s'agit d'un dépouillement peu compliqué et peu étendu. Elle présente, en effet, une série de désavantages qui deviennent surtout sensibles quand le matériel statistique est fort considérable. Ce sont les suivants :

1° Le pointage suppose le maniement fréquent de tout le matériel statistique. Or, ce matériel est pondéreux, encombrant et d'un maniement désagréable, peut-être même dan-

gereux, à cause de la poussière dont il se couvre dans les salles destinées au classement des archives;

2° L'employé peut, en inscrivant la barre, se tromper de colonne;

3° Il peut aussi douter de l'inscription d'une unité;

4° Le contrôle de ces erreurs est généralement impossible dans le cas visé au 2°; il peut être efficace dans le cas prévu au 3°, à la condition que d'autres colonnes soient réservées à des unités de même genre, mais ces corrections exigent beaucoup de temps et une attention scrupuleuse.

162. La seconde méthode est la méthode de comptage; elle porte aussi le nom de méthode des « fiches », par allusion à la forme matérielle sous laquelle elle se réalise.

L'emploi des fiches pour le dépouillement se rattache à l'utilisation du bulletin individuel dans les recensements. Alors que le recensement de la population se faisait encore, en général, au moyen d'évaluation, le comte de Chabrol, préfet de la Seine, donna son adhésion à un projet soumis par ses bureaux, d'après lequel le recensement de Paris se ferait, en 1817, au moyen de bulletins nominatifs. Ces états nominatifs, d'après M. Bloch, s'appliquaient à « chaque location séparée », c'est-à-dire à chaque ménage (1). Le bulletin nominatif par ménage fut étendu à toute la France en 1836, alors qu'à Paris le bulletin collectif avait déjà été remplacé par le bulletin individuel. En Italie, lors du recensement effectué le 31 décembre 1871, le relevé exécuté par les communes avait été transcrit en vue du dépouillement, sur des fiches individuelles portant le même numéro d'ordre que le bulletin (*cartoline*) : l'essai des fiches pour le dépouillement avait déjà été fait en 1867 par Ernst Engel, mais l'expérience faite en Italie lui suggéra l'idée de remplacer le bulletin collectif par un bulletin individuel qui

(1) Cfr. Bloch (M.), *Traité de statistique*, pp. 303 et 359.

équivaut à une fiche. Depuis lors, l'idée du bulletin individuel a fait son chemin; elle se trouve réalisée dans un grand nombre de pays. Dans certains bureaux de statistique, on a préféré s'en tenir au bulletin de ménage, mais on a malgré cela recouru à l'emploi des fiches pour le dépouillement. M. von Mayr a longuement exposé son système caractérissé par l'emploi de couleurs différentes suivant les catégories de fiches et l'emploi de signes ou notations conventionnelles. Lorsque le dépouillement n'est pas centralisé, comme c'est encore le cas en Belgique pour le recensement de la population, les agents des communes chargés des premières opérations doivent employer des fiches pour procéder à leur travail; les instructions sur la matière leur prescrivent la marche à suivre pour utiliser les cartes individuelles et les classer. Il est permis de dire qu'à l'heure actuelle, la méthode de comptage, soit à l'aide de bulletins individuels soit au moyen de fiches, est généralement admise pour les recensements. Comme il s'agit là d'une partie essentielle de la statistique, l'expérience faite dans ce domaine est décisive à l'égard du reste.

163. Dans certains cas, d'ailleurs, l'emploi de la fiche est obligatoire; chaque fois que les données statistiques doivent être extraites d'un registre, il n'y a d'autre moyen d'organiser pratiquement le dépouillement que de transcrire sur fiches les indications à dépouiller; les données se succédant dans un ordre différent du classement statistique, l'agent dépouilleur devrait constamment changer de cadre pour y porter les unités, ce qui serait impraticable.

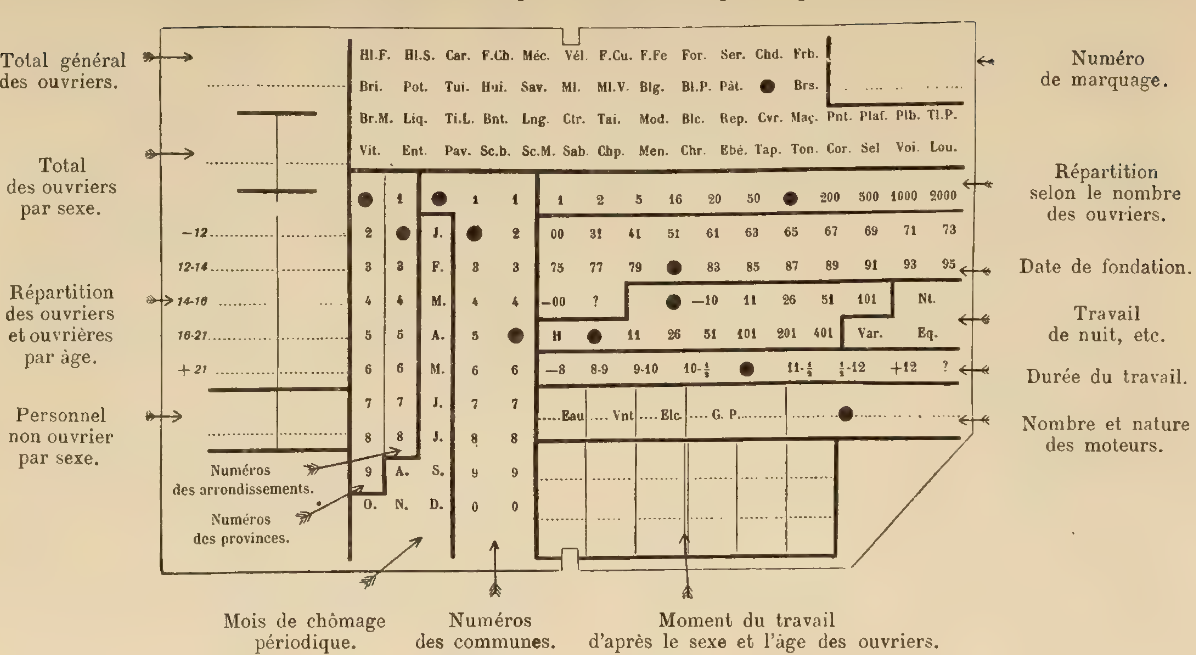
Le pouvoir représentatif de la fiche est beaucoup plus élevé que celui du simple bulletin individuel. Dans un sens large, il est exact de dire que le bulletin individuel est une fiche, mais il ne possède pas, loin de là, toutes les qualités spéciales de celle-ci en vue du dépouillement. Tout d'abord, le bulletin individuel a été manipulé par un grand nombre de personnes, a séjourné dans des locaux plus ou moins

salubres et arrive finalement au bureau central dans un état de propreté qui laisse beaucoup à désirer. Il est encombrant et l'on doit disposer de vastes locaux pour en classer les paquets, qui se remplissent rapidement de poussière. Rien ne distingue les bulletins d'après le sujet dont ils émanent; ni le sexe, ni l'état civil, ni la profession, ni la situation géographique, ni aucune donnée quelconque ne s'y marquent par un signe apparent. Il faut toujours, pour les découvrir, recourir à la lecture, ce qui rend les classements longs et pénibles. Ajoutons que la mauvaise écriture de ces bulletins est un obstacle à leur utilisation rapide et sûre. Aussi, n'est-il pas douteux que la fiche, spécialement préparée en vue du dépouillement, est supérieure au bulletin individuel employé comme fiche.

164. Avant de faire le schéma de la fiche, il faut avoir arrêté les tableaux de présentation. On reporte alors sur les fiches les indications relatives à chaque unité contenues dans les bulletins ou formulaires.

Nous reproduisons en premier lieu le modèle de la fiche employée lors du « recensement général des industries et métiers en Belgique », en 1896. Le type de fiche suivant, utilisé pour la statistique des accidents du travail ayant entraîné une incapacité de travail permanente, combine les indications écrites avec la notation conventionnelle.

Dans la fiche employée lors du recensement général des industries et des métiers en 1896, les indications en abrégé servent à désigner les industries les plus fréquentes; celles qui se rencontrent plus rarement sont indiquées au moyen du numéro de marquage porté à la liste générale. Les annotations typographiques reproduites en marge de la fiche permettront au lecteur de se rendre compte de l'utilisation de celle-ci; ainsi, par exemple, aucun arrondissement administratif ne comprend plus de 199 communes; la commune 125 (classement alphabétique) est indiquée par la perforation des chiffres 1 à la colonne des centaines, 2 à celle des dizaines, 5 à celle des unités.



Modèle de la fiche employée pour le recensement général des industries et métiers en Belgique, en 1896.

N° de l'accident.	Marquage.										N° de l'entreprise.
....	....	1 10	2 11	3 12	4 13	5 14	6 15	7 16	8 17	9 18	....

Nature de l'industrie.	Cause de l'accident.	Nature de la blessure.	N° de l'assureur.
....	....	....	....
Localité.			

Salaire de base.	Durée de l'incapacité temporaire.
....	....

— 15 a. ○	15-20 ○	20-30 ○	30-40 ○	40-45 ○
45-50 ○	50-55 ○	55-60 ○	60-70 ○	70 et + ○

Année de la naissance.			
	43-56 ○	57-91 ○	3-4 m. ○
	5-6 ○	6-9 ○	9-1 a. ○
....	$1\frac{1}{2}$-2 ○	2-$2\frac{1}{2}$ ○	$2\frac{1}{2}$-3 ○

1	0	0
1-14 ○	1	1
	2	2
15-28 ○	3	3
29-42 ○	4	4
	5	5
4-5 ○	6	6
1-$1\frac{1}{2}$ ○	7	7
+ 3 a. ○	8	8
	9	9

—10 % ○	10-20 ○	20-30 ○	30-40 ○	40-50 ○	50-60 ○
60-70 ○	70-80 ○	80-90 ○	90-100 ○	100 % ○	

—10 % ○	10-20 ○	
20-30 ○	30-40 ○	40-50 ○
50-60 ○	60-70 ○	70-80 ○
80-90 ○	90-100 ○	100 % ○

	F.	C.
Allocations journ.	....	....
Frais Méd. et Ph.	....	....
Allocation annuelle (C.).	....	....
Charge provisoire...	....	....

Mort c. r.	Mort o. s. r.	Mort n. cons.	TRANSFORMATIONS P en T	T en P	aggr.	dim.
○	○	○	○	○	○	○

Capitaux pour rentes...	....	....	Pr. ○
Capitaux versés..........	....	....	Pu. ○
Charge finale.........	....	....	

Modèle de la fiche employée pour la statistique des accidents du travail ayant entraîné une incapacité de travail permanente.

165. Les avantages que présente l'emploi des fiches sont nombreux; on pourrait croire que le temps employé à leur confection pourrait être mieux mis à profit; cependant, l'expérience a prouvé que cette méthode est à la fois plus expéditive, plus économique et plus sûre que toute autre :

1° Comme nous l'avons fait déjà remarquer, la méthode des fiches est la seule qui convienne dans le cas où les données statistiques doivent être extraites de registres, comme les registres de l'état civil, par exemple;

2° Le principal avantage des fiches à l'égard du bulletin individuel consiste dans leur pouvoir représentatif, qui est fort étendu. En utilisant des couleurs, des formats et des profils différents, on différencie facilement un grand nombre de données; de cette façon, on supprime des lectures fatigantes et on abrège beaucoup la durée des classements;

3° Un grand nombre de données se trouvant réunies sous un faible volume grâce aux indications conventionnelles portées sur les fiches, il est permis de classer le matériel dans des meubles spéciaux. Les fiches étant réunies par des tringles, on évite d'en égarer, ce qui arrive fréquemment pour les bulletins individuels au cours des nombreux classements qui se font au cours des opérations. De plus, le maniement des fiches n'est pas antihygiénique;

4° Les fiches peuvent être perforées à l'endroit correspondant où telle qualité de l'unité se trouve indiquée. En les groupant d'après tel ou tel caractère, il est facile de s'assurer qu'aucune fiche étrangère à ce classement n'a été, par erreur, glissée dans ce paquet;

5° Les modifications de classe et les combinaisons entre plusieurs données sont rendues beaucoup plus faciles;

6° Après le classement des fiches, les résultats numériques peuvent être obtenus rapidement à l'aide des machines à calculer, sans transcription nouvelle des données.

CHAPITRE III

Le dépouillement statistique et le calcul par les machines

I. — Machines à dépouiller.

166. Comme il était naturel, l'homme s'est appliqué, depuis longtemps, à trouver des dispositifs capables d'atténuer ou de supprimer le travail fatigant du calcul. Le lecteur curieux de détails trouvera, sur cette partie intéressante des sciences appliquées, dans les ouvrages spéciaux la description complète de ces appareils (1).

L'aperçu le plus sommaire que nous en donnerions ici nous éloignerait de notre matière : nous n'avons qu'à dire, en peu de mots, en quoi consistent les principaux appareils à dépouiller et à calculer qui sont en usage dans les bureaux de statistique et à en montrer les avantages.

Le machinisme spécial de la statistique comprend deux catégories d'appareils : ceux à calculer et ceux à dépouiller. Nous parlerons d'abord des seconds et, parmi eux, nous choisirons les deux dispositifs les plus typiques et les plus usuels : la machine électrique à dépouiller du D[r] Hermann Hollerith et le classi-compteur-imprimeur de M. Lucien March (2). Toutefois, avant d'en aborder la description, nous avons encore à indiquer quelques considérations générales.

M. Hollerith a fait connaître lui-même à la Société de

(1) Cfr. Maurice D'OCAGNE : *Le calcul simplifié par les appareils mécaniques et graphiques*, Paris, 1905. — L. JACOB : *Le calcul mécanique; appareils arithmétiques et algébriques intégrateurs*, Paris, 1911. — Maurice D'OCAGNE : *Calcul graphique et nomographie*, Paris, 1914.

(2) Dans la division de la matière, nous n'avons pas suivi l'ordre adopté par certains auteurs, par exemple von Mayr, qui distingue deux classes parmi les appareils selon qu'ils sont actionnés mécaniquement ou à la main. Ce principe de distinction n'a rien d'essentiel et, en fait, les appareils à la main se transforment très souvent en appareils mécaniques.

statistique de Londres, en 1894, comment, ayant pris part en qualité d'agent spécial au recensement des Etats-Unis en 1880, son attention fut attirée sur la nécessité de trouver quelque dispositif mécanique permettant de simplifier le travail de dépouillement et surtout de l'intensifier. Les classements successifs auxquels il faut procéder pour pouvoir dépouiller les bulletins sous différents aspects prennent un temps énorme dans les procédés ordinaires de dépouillement, particulièrement quand il s'agit de bulletins de ménage ; bien qu'il y ait progrès sous ce rapport quand on a recours au bulletin individuel, le travail de classement n'en demeure pas moins une gêne ; reprenant les liasses de bulletins, les agents du relevé doivent feuilleter rapidement des centaines de milliers de documents pour en retirer un certain nombre ou opérer un classement. L'état de malpropreté de ces bulletins rend le travail aussi peu agréable que possible et la variété des écritures, généralement mauvaises, amène bientôt une réelle fatigue visuelle. Or, ces classements successifs sont indispensables dans les travaux de statistique. C'est par la multiplicité de ses travaux que la statistique peut arriver à dégager les caractères essentiels des phénomènes collectifs complexes. Dans les procédés anciens, le statisticien doit souvent sacrifier l'une ou l'autre combinaison de données parce qu'elle exigerait trop de travail et trop de temps. Au fond, la question du machinisme n'est pas seulement une question d'ordre technique ou pratique : c'est une question scientifique à laquelle est liée, dans certains cas, la réalisation du but supérieur de la statistique.

167. La machine électrique de M. Hollerith permet la solution des trois problèmes statistiques suivants : 1° dépouiller et additionner chaque classe séparément ; 2° séparer les fiches selon telle ou telle catégorie ou selon des points de vue donnés ; 3° additionner les résultats des combinaisons de diverses catégories.

Le dispositif Hollerith comporte trois parties distinctes qui correspondent à autant d'opérations différentes : l'appareil à perforer; la machine à classer; le mécanisme permettant de lire, de combiner et d'additionner les fiches. Nous en donnerons une description sommaire, l'essentiel n'étant pas tant, au point de vue auquel se place cet ouvrage, la précision technique que les résultats réalisés à l'aide de ce mécanisme.

L'utilisation de la machine suppose tout d'abord que l'on a confectionné autant de fiches qu'il y a de bulletins individuels ou de données individuelles sur les bulletins de ménage. La fiche dont on se sert est en carton; elle porte des indications conventionnelles correspondant aux diverses qualités et nombres que peuvent présenter les unités à dépouiller. Généralement, ces indications sont réalisées par des combinaisons de chiffres : par exemple, les provinces, arrondissements, communes d'un pays, sont désignés au moyen d'un numéro d'ordre; ce numéro se marque sur la fiche en réservant, à chacune de ces divisions administratives, autant de rangées verticales de neuf chiffres que le numéro d'ordre comporte de rangées et en pointant, dans chaque colonne, successivement le chiffre qui correspond aux unités, dizaines, centaines. Pour d'autres données, on se sert simplement des initiales du mot qui les désigne, comme : C = célibataire, M = marié, V = veuf, D = divorcé, S = séparé de corps. Grâce à ces combinaisons, on arrive aisément à porter sur une fiche de dimensions restreintes (0,15 × 0,08) un très grand nombre de données. Voici, par exemple, le modèle de la fiche employée pour le recensement de l'agriculture aux Etats-Unis, en 1900. Les indications imprimées « en clair » ont été ajoutées pour la facilité du lecteur : elles ne figurent pas sur le modèle original. En peu de temps, les employés sont familiarisés avec les signes conventionnels et avec la place qu'ils occupent sur la carte.

I 3 5 7 2 4 6 8	State.	1 3 5 7 2 4 6 8	1 3 2 4	1 3 2 4	County.	x	5 7 6 8	1 3 5 7 2 4 6 8	x	W B Race. Ch In	x	Ow PO OT Mg CT ST Tenure.	x		x

Size of Farm.	Source of Income.	Amount of Income	Area.			Improved Area.			Total Value of Farm				Value of Buildings.			Value of Imple-ments.			Value of Products			Value of Products Fed to Live Stock.			Value of Live Stock.			Amount Expended for Fertilizers.			Amount Expended for Labor.		
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

Modèle de la fiche employée pour le recensement de l'agriculture aux États-Unis, en 1900.

168. La carte perforée peut être assimilée, comme l'a dit M. Cheysson, au carton du métier Jacquard, mais au lieu de fils réglant les mouvements des navettes, la carte laisse passer un courant électrique. La première besogne consiste donc à transformer les bulletins individuels en cartes perforées. Cette partie du travail présente une grande importance. Elle s'exécute à l'aide de machines à perforer dont il existe plusieurs types. Le plus ancien porte le nom de Keyboard-Punch; en plaçant une pointe indicatrice sur une feuille métallique reproduisant les indications de la fiche,

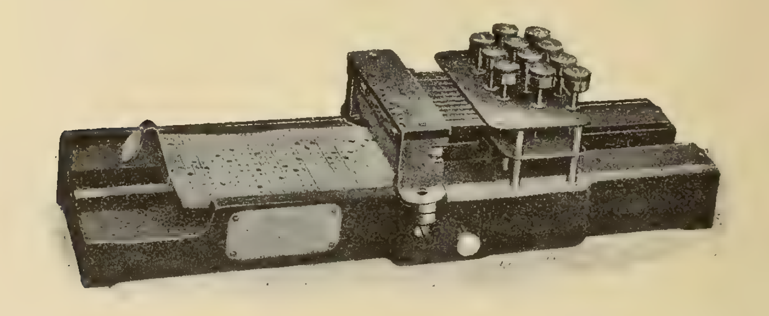


Fig. 1.

l'opérateur, au moyen d'un dispositif analogue à celui du pantographe, perfore la fiche elle-même à l'endroit correspondant. Le Keyboard-Punch est assez lent, attendu qu'il ne peut perforer qu'une seule fiche à la fois. M. Hollerith, lors de la IXe session de l'Institut international de statistique, en 1895, annonçait déjà que ce premier appareil serait remplacé par un autre plus rapide. Le perfectionnement annoncé a été réalisé au moyen d'un appareil (Key-Punch) dont l'aspect général est indiqué au moyen de la *figure 1* ci-dessus.

Le Key-Punch est un appareil plus simple que le premier. Il consiste en une solide boîte métallique pourvue de onze poinçons dont dix numérotés de 0 à 9, plus un poinçon désigné par la lettre x. La carte placée sur une plaque mobile

est insérée par l'opérateur dans l'appareil jusqu'à ce que les quatre premières colonnes imprimées à partir du côté gauche soient placées sous les poinçons; on suppose que toutes les notations conventionnelles de la fiche sont exprimées par des chiffres, comme sur le modèle de la fiche employée pour le recensement agricole. La lettre x est imprimée dans les cases de la fiche correspondant à la donnée « néant » du bulletin. Elle sert à indiquer qu'il n'y a aucune donnée à dépouiller relativement à ce point et, en attirant forcément l'attention de l'employé sur chaque case de la fiche, elle évite des erreurs tout en facilitant le contrôle.

A part les cases où figure la lettre x, on doit faire une perforation à chaque colonne verticale de la fiche : une ferme de 12 acres sera indiquée en poinçonnant aux endroits 0, 1, 2 de la case réservée à l'étendue cultivée.

Le Key-Punch se construit actuellement d'après un modèle actionné par l'énergie électrique; le rendement de cette machine à perforer est très considérable. En général, ce travail est confié à du personnel féminin.

169. Le Gang-Punch est une machine plus robuste, capable de perforer plusieurs cartes à la fois. Certaines indications se répétant plusieurs milliers de fois — comme le nom des provinces et des villes — on peut augmenter le rendement de la machine en lui faisant perforer une dizaine de fiches à la fois. Beaucoup de données se répètent même dans les bulletins faisant partie des dépouillements les plus compliqués.

Dans un recensement industriel, par exemple, on trouve des milliers de bulletins se rapportant à des personnes du sexe masculin exerçant un même métier, exploitant une entreprise individuelle, n'employant que des membres de leur famille. Des bulletins de l'espèce peuvent être aisément traduits en fiches par séries, à l'aide du Gang-Punch, dont on verra à la page suivante une reproduction (*fig. 2*).

Les résultats atteints au moyen des machines à perforer

seront indiqués et critiqués au moment où nous entreprendrons l'étude du rendement et des avantages du machinisme statistique. En attendant, nous continuons la description de l'appareil en nous efforçant de la rendre aussi intelligible que possible.

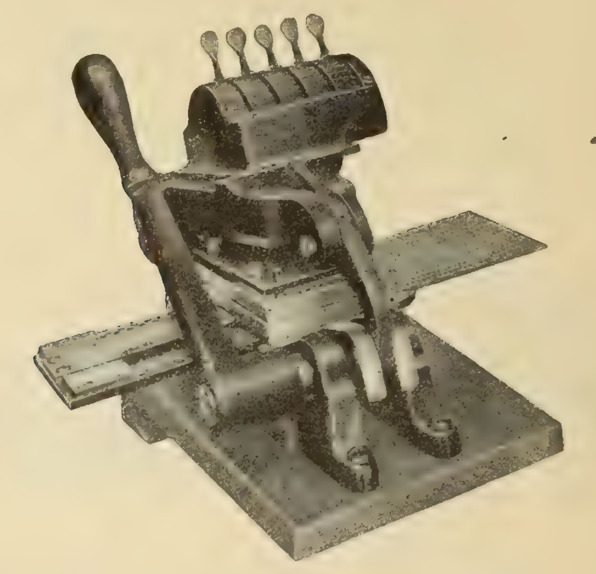


Fig. 2.

170. Une des difficultés les plus ardues du dépouillement consiste dans le classement des bulletins d'après les points de vue nombreux que requiert la présentation en tableaux. De toute façon, il faut employer des fiches; le classement des bulletins individuels, qui peuvent être assimilés à des fiches, reste fort lent et pénible. M. Hollerith s'est donc demandé s'il n'existait pas quelque moyen de classer automatiquement les fiches statistiques perforées et il y est arrivé par l'ingénieux appareil dont la *figure 3* (p. 270) représente l'aspect d'ensemble. M. le D[r] Jacques Bertillon en a donné une description très claire (1) que nous utilisons dans le résumé qui suit.

---

(1) Dans *La Nature* du 17 janvier 1914.

Sur le plateau supérieur de la machine délimité par le bâti en fonte, se placent les cartes perforées, mises verticalement dans le sens de l'impression. Les fiches sont réunies par paquets et sont serrées au moyen d'une barre entraînée par des poids, de sorte que le paquet est toujours compact. La première carte se trouve au-dessus d'une fente où elle s'engage et elle y est poussée par un cadre mobile (partie supérieure du bâti) qui monte et descend alternativement et appuye sur la première fiche qui se présente en dessous de lui.

La carte se trouvant poussée sous la plate-forme s'engage entre des rouleaux qu'on aperçoit sur la *figure 3*; celui d'arrière est en ébonite, celui d'avant est en cuivre; ce dernier est chargé d'électricité par une languette : la source d'électricité est le petit moteur placé au pied de l'appareil.

S'appuyant sur le rouleau en ébonite, une forte aiguille métallique vient reposer son extrémité, munie d'une pointe verticale, sur le rouleau en cuivre, ou plutôt sur la carte qui s'engage entre les rouleaux. L'aiguille est mobile dans le sens du rouleau, de sorte qu'on peut faire promener son extrémité sur celle des colonnes de la fiche qui correspond à l'unité d'après laquelle les cartes doivent être classées. S'agit-il, par exemple, de classer les exploitations agricoles de 100 acres et plus (-100, 100-200, 200-300, etc.), l'aiguille sera placée en regard de la quatrième ligne verticale de la fiche reproduite page 265. Aussi longtemps que l'aiguille se promènera sur une surface pleine, il ne se passe rien, mais quand l'aiguille rencontre un trou perforé dans la carte, son extrémité entre en contact avec le rouleau chargé d'électricité et aussitôt la carte sera entraînée vers un des godets classeurs, placés sous la plate-forme, au nombre de onze : un pour les zéros, 9 pour 1 à 9, un pour les erreurs.

La description du mécanisme servant à diriger la carte vers le godet qui lui est réservé et non vers un autre est

d'essence trop technique pour pouvoir être abordée ici. On en trouvera une dans l'article déjà cité de M. le D[r] Bertillon.

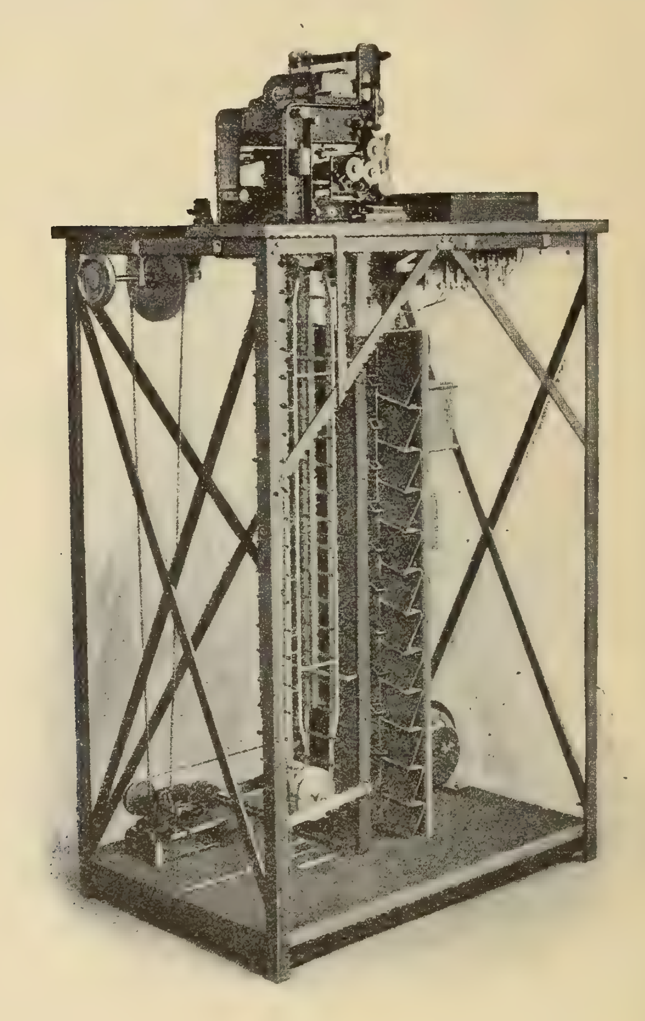


Fig. 3.

171. Il ne suffit pas de classer les fiches, il faut surtout les dépouiller, enregistrer les unités qui y sont marquées à l'aide de perforations et ajouter les unes aux autres celles qui sont de même nature. C'est ce que réalise la

troisième et la plus importante partie de la machine, composée de trois parties essentielles : *a*) l'appareil de contact; *b*) un jeu de compteurs électriques; *c*) le système des relais.

L'appareil de contact est peut-être la partie la plus neuve du système imaginé par M. Hollerith. Il comprend un plateau fixe posé sur la table, horizontalement, et un plateau mobile se rabattant sur le premier à l'aide d'un levier manœuvré à la main par une poignée. Le plateau mobile contient dans son épaisseur une série d'aiguilles métalliques symétriquement disposées, attachées à un ressort en boudin suffisamment distendu pour que les extrémités des aiguilles sortent quelque peu de la boîte où elles sont renfermées. Le plateau fixe présente d'abord une plaque métallique, de la grandeur de la fiche à employer, percée d'autant de trous qu'il y a d'endroits à perforer indiqués sur la fiche; ces trous, les aiguilles et la fiche correspondent exactement. Quand on met une fiche perforée sur le plateau fixe, les trous de la fiche correspondent aux trous de la plaque métallique et au-dessous on voit se montrer autant de petits godets remplis de mercure (1) (*fig. 4*). Qu'on abaisse, sur la fiche perforée, le plateau mobile, les aiguilles rencontreront le papier en certains endroits et se replieront à l'intérieur, grâce au ressort à boudin auquel elles sont attachées; les autres ne rencontrant aucun obstacle plongeront, à travers les trous perforés, dans les godets de mercure; un courant opposé étant rattaché à chaque plateau, le circuit se trouve fermé par l'introduction des aiguilles dans les godets, mais seulement pour les données correspondant à ces récipients de mercure.

Dans la machine employée au dépouillement du recensement agricole, il n'y avait pas moins de 354 aiguilles, et, par conséquent, autant d'ouvertures dans la plaque métal-

---

(1) Voyez page 272 (*fig. 4*), une reproduction de la machine complète: l'appareil de contact est placé sur la table, à droite.

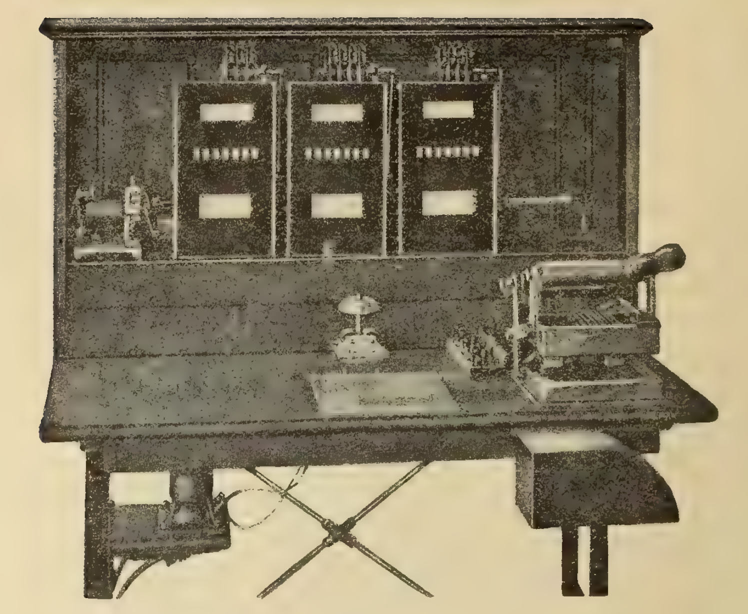


Fig 4.

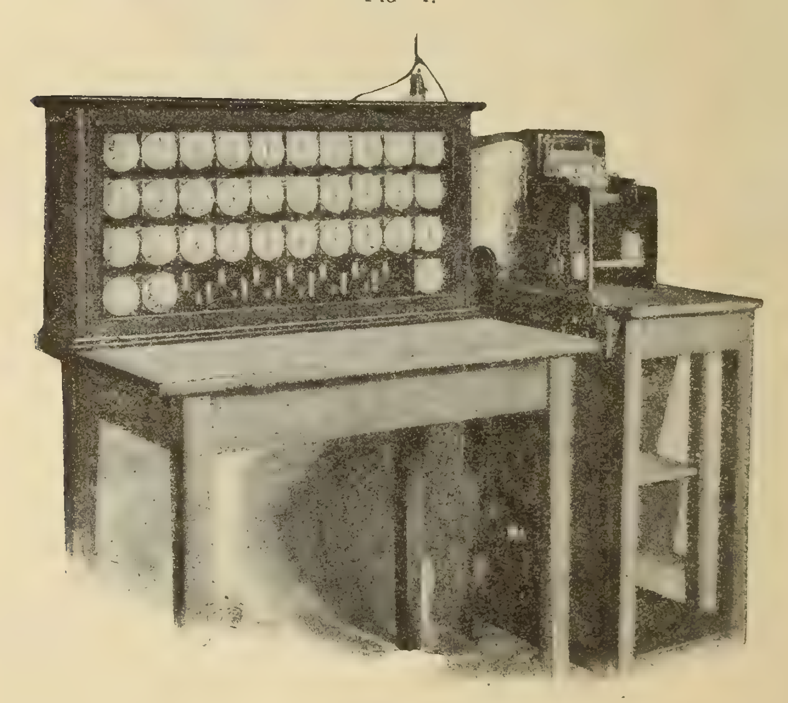


Fig. 5

lique et de godets. A chaque godet correspond un compteur électrique, composé de quatre chiffres et qui avance d'une unité à chaque contact. Après un certain temps, on relève les chiffres indiqués par les compteurs et on remet ceux-ci à zéro. Dans le plan primitif, le travail des compteurs était rendu apparent au moyen de cadrans pourvus d'une aiguille mobile (*fig.* 5). Actuellement, la disposition est

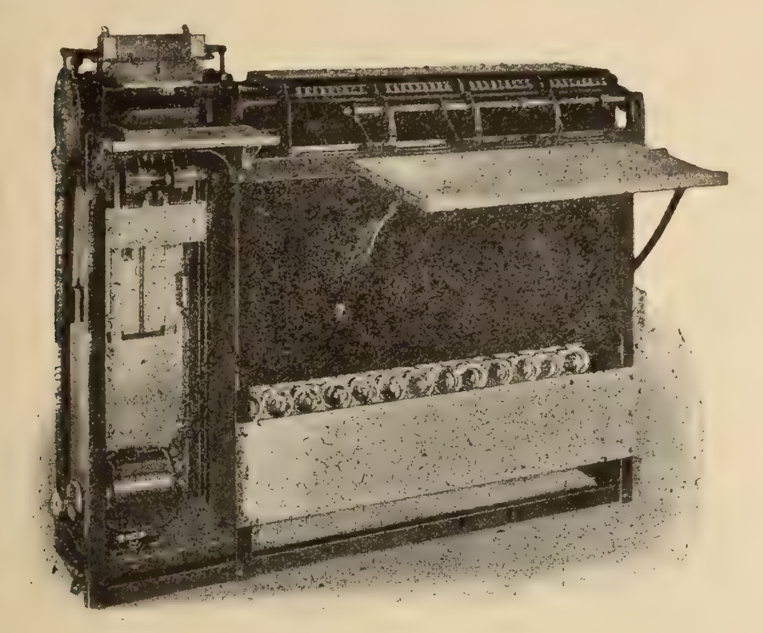


Fig. 6.

différente; les chiffres apparaissent directement dans les lucarnes et les fiches sont entraînées automatiquement vers des boîtes réservées chacune au comptage d'un fait particulier (*fig.* 6). Le comptage effectué de la sorte répond à toutes les nécessités des tableaux primaires dans lesquels les unités sont comptées séparément : tant d'hommes, de femmes, de patrons, d'employés, d'ouvriers. La machine réalise ainsi les desiderata ordinaires de la statistique, mais lorsqu'on analyse les résultats d'un relevé, on a souvent besoin de réunir des données plus compliquées. Ainsi,

veut-on présenter séparément les données relatives aux établissements industriels qui employent des ouvriers étrangers à la famille des exploitants et utilisent une force motrice, ou si l'on désire savoir combien d'enfants sont nés après dix ans, du mariage d'un père de tel âge et d'une mère de tel autre, on se trouve en présence de combinaisons qui ne peuvent se réaliser qu'à l'aide de classements répétés, longs et fastidieux. En dehors du système des fiches, de telles combinaisons de données sont même radicalement impossibles; elles ne sont guère plus réalisables pratiquement lorsqu'on dispose de bulletins individuels. Seule, la fiche perforée, dont on peut vérifier le classement à l'aide d'une tige métallique traversant, sans rencontrer d'obstacle, tous les trous correspondant à ces indications, est capable d'assurer le dépouillement de semblables données.

La machine Hollerith assure très facilement la solution de ce problème à l'aide des « relais » ou connexions électriques dont elle est pourvue. Au début, ces combinaisons ne pouvaient se faire qu'à la condition d'un assez long travail de préparation, et chaque problème statistique se doublait d'un problème électrique exigeant la coopération d'un technicien. Mais, dès 1895 déjà, d'importants perfectionnements ont été apportés à l'appareil; l'adjonction d'un commutateur général a eu pour effet, d'après M. Rauchberg, « de faciliter par avance toutes les combinaisons possibles des diverses parties essentielles de la machine, en les mettant les unes avec les autres dans une relation telle qu'une seule cheville ou deux chevilles reliées par un fil conducteur suffisent pour ramener les rattachements et les transmissions de courant désirables ». On trouvera, dans le rapport que M. Rauchberg adressa en 1895 à l'Institut international de statistique (1), un aperçu général

(1) *Bulletin de l'Institut international de statistique,* t. IX, première livraison, p. 249. Rome, 1895.

du fonctionnement du commutateur général; ces détails nous semblent en dehors du cadre d'une description non technique, telle que la nôtre, et nous nous bornons à renvoyer le lecteur aux mémoires techniques relatifs au fonctionnement de la machine. Nous devons noter toutefois un perfectionnement important apporté à la machine Hollerith par le service du *Census*. Ce service, qui construit lui-même les appareils dont il se sert, a imaginé un dispositif permettant d'imprimer directement les données au compteur, au lieu de les copier; la compagnie qui exploite les brevets Hollerith annonce qu'elle appliquera prochainement à ses machines un appareil de l'espèce.

172. Le principe de marquage, de classification et de tabulation des machines Hollerith a été lui-même l'objet de plusieurs perfectionnements dans le pays où il a été trouvé et appliqué pour la première fois : les Etats-Unis d'Amérique. Ainsi, les machines Hollerith, jusqu'à présent du moins, n'enregistrent pas elles-mêmes les données sélectionnées qu'elles ont pour objet de compter; c'est à un opérateur travaillant à la main, qu'incombe ce soin, et pour ce faire l'opérateur est tenu d'arrêter la machine, puis de la remettre en marche. Les machines « Powers » ont réalisé un perfectionnement important sous le rapport de la sélection des cartes et surtout sous celui de la transcription des données qui se fait automatiquement; après avoir été sélectionnées, les cartes sont portées à la machine tabulatrice qui enregistre en clair et sous forme de tableaux les renseignements et totalise automatiquement les chiffres des colonnes; cependant l'opération exige encore le transport à la main des fiches sélectionnées jusqu'à la machine à tabuler. De nouveaux appareils, les machines « Pierce », que nous avons vues fonctionner en Amérique, présentent de précieux avantages. Ainsi, les fiches peuvent contenir indifféremment des lettres et des

chiffres, ce qui augmente beaucoup le pouvoir de représentation des cartes; la perforation se fait à l'aide d'une machine, pourvue d'un clavier semblable à celui d'une machine à écrire; le fonctionnement de l'appareil perforateur est à la fois rapide, sûr et assez doux pour être confié à un personnel féminin. La sélection des cartes se fait par des machines à classer opérant dans le sens horizontal, au lieu du sens vertical comme dans les machines Hollerith; ces appareils conduisent automatiquement les cartes sélectionnées jusqu'à l'appareil tabulateur. Celui-ci reproduit en clair tous les renseignements, sous forme de tableaux et en fait l'addition par colonnes de chiffres. Cet ensemble de dispositifs représente une somme importante de perfectionnements sous le rapport de la rapidité et de l'épargne de main-d'œuvre.

173. Le classi-compteur-imprimeur de M. Lucien March est un appareil tout à fait différent de la machine électrique de M. Hollerith. Beaucoup plus simple, il n'en rend pas moins de grands services dans le travail de dépouillement et paraît même mieux adapté à la besogne journalière d'un bureau de statistique. Le principe et le fonctionnement du classi-compteur-imprimeur ont été exposés par M. Georges Vitoux, dans un article paru dans *La Nature* (11 mai 1901, p. 369) auquel nous nous référons.

Le problème dont M. March a cherché la solution, est triple : enregistrer simultanément les divers aspects sous lesquels une unité statistique est envisagée dans les cadres (sexe, état civil, âge, profession, lieu de naissance d'un individu); supprimer le travail du relevé des compteurs, ou plutôt permettre l'enregistrement simultané et rapide des indications qu'ils portent; en même temps, trouver un dispositif de contrôle d'après lequel l'exactitude du travail effectué par l'employé pourrait être appréciée.

L'appareil n'est pas très volumineux: deux fois les dimensions d'une machine à écrire. Le modèle que nous avons

sous les yeux (*fig. 7*) et qui est utilisé dans nos bureaux est assez lourd; il est probable que depuis l'époque déjà éloignée de son acquisition, les appareils nouveaux ont dû perdre quelque chose de cette apparence massive.

La partie antérieure est une tablette légèrement inclinée comme celle d'une machine à additionner, comprenant une soixantaine de touches formant un clavier. Ces touches ne portent aucune désignation, il appartient au statisticien de décider, lors de la mise en train, à quelle donnée correspondra telle touche de l'appareil. A cet effet, une petite ouverture ménagée sous un bouton en matière transparente permet de glisser une inscription manuscrite, chiffre ou lettre conventionnelle ou « en clair ». Nous avons vu remplacer ce système par des inscriptions portées directement sur la tablette à côté ou en dessous des touches; les opérateurs nous ont aussi fait remarquer que le toucher serait plus agréable à l'aide de touches en ébonite, légèrement creusées pour recevoir l'extrémité du doigt.



Fig. 7.

Chaque touche correspond à un compteur formé de quatre chiffres. Lorsqu'on abaisse une touche, celle-ci vient appuyer sur un levier qui oblige, par l'intermédiaire d'une tige de commande dont elle dépend, le compteur auquel elle se relie à avancer d'une division. L'opérateur abaisse successivement toutes les touches qui correspondent aux diverses indications de la fiche à dépouiller, puis, en abaissant la manette placée à droite de la machine, il fait avancer d'un rang tous les compteurs intéressés. Il continue de la sorte jusqu'à la fin de la série à dépouiller.

Les chiffres marqués par les compteurs apparaissent, à l'arrière de la machine, par une lucarne ménagée dans la plaque de cuivre qui recouvre l'appareil. Lorsqu'on veut totaliser, on étend sur les compteurs un ruban chargé d'encre grasse et on rabat le cadre mobile situé à l'arrière. Ce cadre est formé de six rouleaux, correspondant aux rangées de compteurs; sur ces rouleaux s'enroule une feuille de papier. Il suffit d'une légère pression pour imprimer les caractères apparents des compteurs. Le papier est ensuite enroulé légèrement et l'appareil, la partie mobile relevée, est prêt pour recevoir de nouvelles inscriptions après qu'on a remis les compteurs au zéro en donnant un tour de la manivelle située en arrière du levier. L'appareil de contrôle est placé dans le socle de la machine : il consiste en un compteur spécial qui enregistre le numéro de l'opération sur une bande de papier se déroulant automatiquement. En regard, des perforations, exécutées par des aiguilles actionnées par les touches, correspondent à la place des compteurs mis en action. Il suffit de confronter ces indications aux mentions du bulletin pour s'assurer si celui-ci a été correctement dépouillé; bien entendu, ce contrôle ne se fait que cinq ou six fois pour cent bulletins dépouillés, c'est-à-dire que l'on juge le dépouillement, sous le rapport de son exactitude, par voie d'essai.

II. — Machines à calculer (1).

174. Les machines à calculer sont, à la fois, beaucoup plus anciennes et plus nombreuses que les machines à dépouiller. Ces dernières sont nées le jour où l'étendue et la complication du matériel statistique sont devenues telles qu'elles dépassaient les ressources en personnel et en argent des bureaux de statistique ; ce fait est de date récente, puisqu'il résulte du dernier état de développement des recherches statistiques. Au contraire, le besoin d'une aide matérielle dans le pénible travail du calcul a toujours été vivement ressenti par les mathématiciens, les astronomes, etc., dont les travaux sont bien antérieurs à ceux des statisticiens. Les plus grands noms des mathématiques se retrouvent dans la liste des inventeurs de machines à calculer : en 1617, Neper indiquait dans sa *Rhabdologie,* publiée à Edimbourg, le procédé connu sous le nom des « bâtonnets de Neper », à l'aide duquel on peut trouver immédiatement le produit d'un facteur multiplié par un chiffre ; en 1642, Blaise Pascal inventa la machine à calculer, véritable machine arithmétique, la première réalisée dans ce genre, qui fut imitée et améliorée un grand nombre de fois ; Leibnitz, l'illustre fondateur du calcul différentiel, imagina en 1671 et parvint à réaliser, en 1694, sous une forme imparfaite encore, une machine à multiplier par additions successives.

Les appareils à calculer usités dans les bureaux de statistique sont ceux qui servent à effectuer les opérations élémentaires de l'arithmétique. Nous en passerons en revue quelques types, non sans avoir prévenu le lecteur qu'il existe d'autres modèles que ceux décrits et que l'auteur n'a aucunement intention de recommander une marque plu-

---

(1) La partie historique et scientifique de cet exposé est faite d'après l'ouvrage de M. Maurice d'Ocagne : *Le calcul simplifié par les procédés mécaniques et graphiques,* Paris, 1905.

tôt qu'une autre, les modèles cités l'étant uniquement à titre exemplatif.

175. M. d'Ocagne appelle « instruments arithmétiques » les appareils qui permettent d'effectuer manuellement les

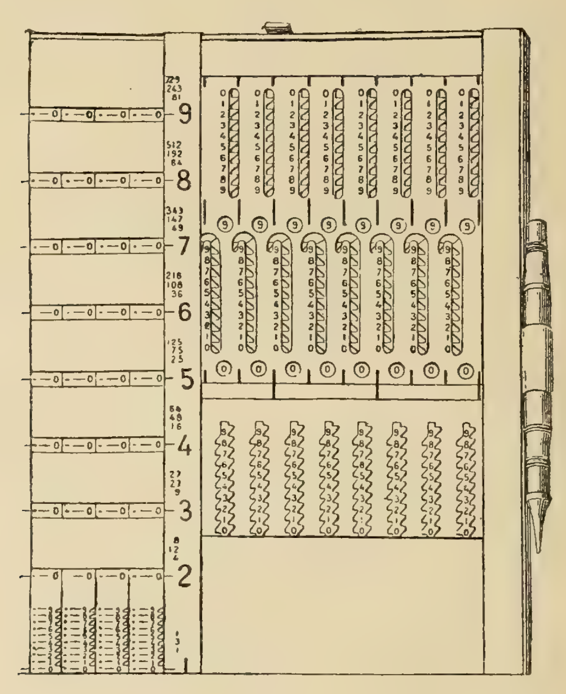


Fig. 8.

opérations de l'arithmétique sans le secours d'aucun mécanisme : ressorts, cames, engrenages, etc.

Au nombre de ces appareils, figure l'arithmographe « Troncet », qui se compose d'une série de colonnes recourbées en forme de crosse, comme dans le modèle ci-dessus (*fig. 8*). En regard de chaque colonne sont écrits les chiffres 0 à 9 et en face on voit un creux; si les dents qui comprennent ce creux sont blanches, on parcourt la fente longitudinale, à l'aide d'un stylet, de haut en bas, jusqu'au butoir inférieur; si elles sont noires ou rouges, on opère

le mouvement inverse, jusqu'à l'extrémité recourbée en forme de crosse.

Soit le nombre 6.478 à faire entrer dans une addition ; il n'y a qu'à placer la pointe mousse du crayon successivement à côté des chiffres 6, 4, 7, 8 pour que le crayon lui-même, par son propre mouvement sur les lames chiffrées, fasse apparaître le total de l'addition.

L'arithmographe permet d'effectuer les quatre opérations fondamentales de l'arithmétique. Le bas prix de l'appareil, dont le modèle le plus complet (opérant sur les nombres de 12 chiffres, pourvu d'un effaceur et d'un multiplicateur à 8 chiffres) ne coûtait (avant la guerre) que 40 francs, est une caractéristique qui mérite d'être signalée.

176. Un autre additionneur, d'un type imité de la machine Pascal, a été réalisé en 1841 par le Dr Roth. « Dans l'additionneur Roth, dit M. d'Ocagne, la chiffraison, au lieu d'être inscrite sur la surface latérale d'un cylindre, l'est le long d'un disque qu'on fait tourner directement au moyen d'une pointe introduite entre les dents fixées à sa tranche. La chiffraison de sens contraire correspondant à la soustraction est marquée en rouge sur un cercle concentrique au premier, chacune de ces chiffraisons paraissant à une lucarne distincte (*fig. 9*).

« Le perfectionnement mécanique réalisé par rapport à la machine de Pascal et à toutes celles qui en sont dérivées, tient à ce que les appareils de retenue, d'un système d'ailleurs entièrement nouveau, sont disposés de façon à n'avoir jamais à fonctionner simultanément » (1).

La machine de Roth a été imitée et perfectionnée. Dans un appareil que nous avons sous les yeux, nous voyons qu'on peut additionner de droite à gauche, ou de gauche à droite, dans le sens de l'écriture et de l'énonciation des

(1) Cfr. D'OCAGNE, *loc. cit.*, p. 32.

nombres à la dictée. Les chiffres de contrôle marquent toujours le dernier enregistrement; le résultat apparaît en chiffres rouges dans les lucarnes supérieures. Après chaque enregistrement, on remet les cadrans à zéro en poussant sur le déclic, à gauche. L'effaceur du résultat total agit très rapidement : il suffit d'introduire le stylet dans la fente et de pousser quelques coups vers la gauche jusqu'à ce que les lucarnes supérieures ne laissent plus apercevoir que des zéros. L'opération est un peu plus compliquée pour la sous-

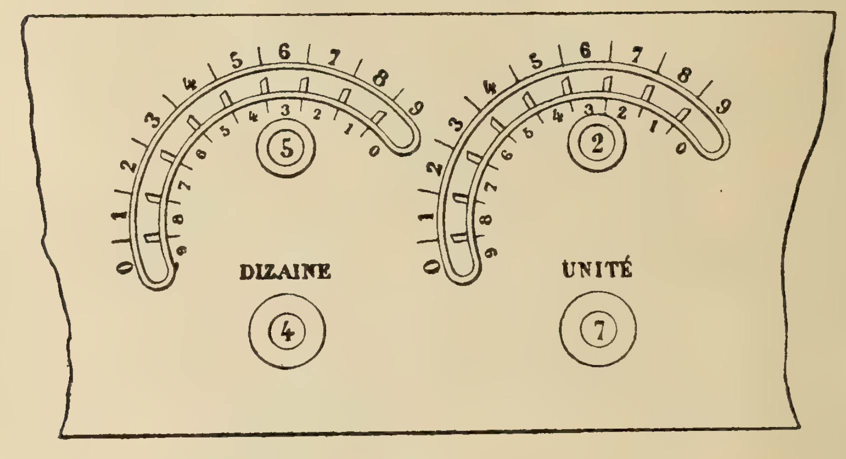


Fig. 9.

traction; elle doit se faire en commençant vers la droite (unités).

177. Les machines à additionner, à touches, paraissent avoir eu pour premier objectif de simplifier la notation des nombres. Elles ont été l'objet d'une véritable exploitation industrielle en Amérique, où de nombreux perfectionnements leur ont été apportés. Les machines à additionner-imprimer datent de 1888; certains types sont bien connus en France et en Belgique, où ils sont employés dans tous les grands établissements financiers. La machine à additionner, à touches, est composée d'un clavier, légèrement incliné, comprenant pour chaque ordre décimal une colonne

de neuf touches numérotées de 1 à 9, la rangée des 1 occupant la place inférieure, puis la rangée des 2, et ainsi de suite. Les machines diffèrent d'après le nombre de rangées; les plus grandes en ont jusqu'à 12 ou 14, les plus usuelles en ont 9; le clavier est alternativement composé de rangées verticales noires et blanches, les deux premières — noires

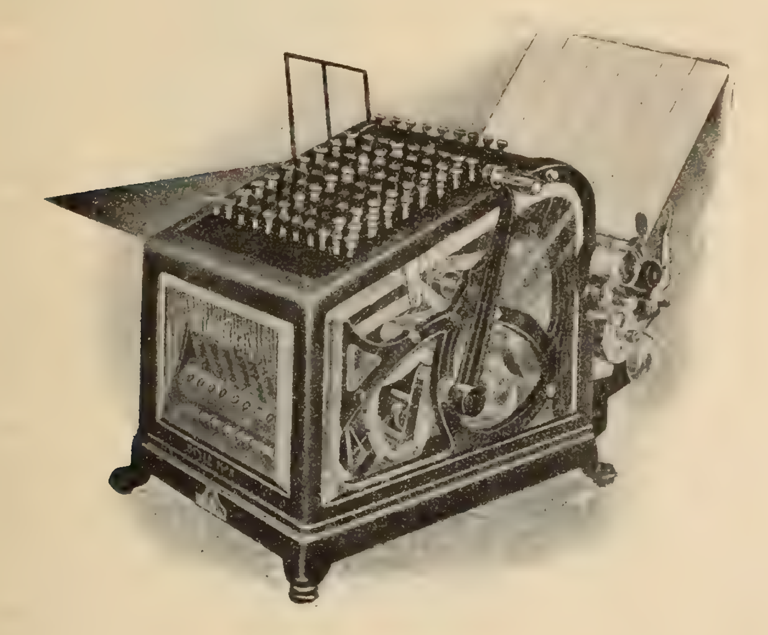


Fig. 10.

— (à droite) correspondant aux centièmes, puis les trois suivantes — blanches — aux unités, centaines et milliers, les trois suivantes — noires — aux unités, dizaines et centaines de l'ordre des mille, puis la dernière — blanche — aux millions. Nous reproduisons ici un des types les plus connus, la machine « Burroughs », imprimant son travail sur rouleau ou sur feuille (*fig. 10*).

Une quantité d'améliorations successives ont été apportées aux machines à additionner. Ainsi, elles possèdent aujourd'hui des touches permettant de corriger, rangée par

rangée, les inscriptions erronées et une touche effaçant d'un seul coup l'inscription tout entière ; elles sont aussi munies d'un bouton qu'il suffit de presser pour « éliminer » un chiffre qu'on veut imprimer sans le comprendre dans l'addition et un autre bouton permettant de répéter un facteur sans le « frapper » chaque fois. Le total apparaît dans des lucarnes, sans qu'il soit besoin de faire agir le mécanisme imprimeur du total. Mais l'amélioration la plus heureuse a été celle qui consista à substituer la manœuvre électrique à la marche à la main. Une réglette, fixée à droite, permet à l'opérateur, par le plus léger attouchement, d'enregistrer et de totaliser les données. La fatigue est entièrement supprimée, grâce à l'électricité, qui est fournie par un petit moteur fixé sous la machine. Le rouleau conducteur du papier est aussi très ingénieusement combiné.

Les machines additionnant et imprimant à la fois sont des plus utiles dans les bureaux de statistique où l'addition joue un grand rôle dans les travaux les plus usuels. Quant au dispositif imprimeur, il a pour les bureaux de statistique une importance de premier ordre en ce qu'il permet de contrôler le travail des employés.

Certaines variétés de la machine à touches sont à écriture visible. Ce détail peut avoir une influence heureuse sur le travail d'un débutant, mais il ne paraît pas maintenir cet avantage quand il s'agit d'un employé entraîné. L'emploi de l'électricité et l'impression des données confiées à la machine sont les premières qualités qu'on recherchera dans un bureau de statistique : la première, à raison de la facilité et de la rapidité du travail, la seconde, parce qu'elle place l'employé sous un contrôle permanent, qui est nécessaire dans les travaux de l'espèce.

Pour éviter la multiplicité des touches, on a imaginé la machine (à écriture visible) comprenant onze touches seulement : pour inscrire un nombre, il suffit d'appuyer sur les touches, en prenant les chiffres dans l'ordre de la dictée ou de la lecture, sans omettre de marquer aucune tranche,

même si elle est formée d'un zéro, ni aucune décimale si la précision du calcul va jusque-là. La disposition du clavier, très ingénieuse, permet à l'opérateur d'acquérir une grande dextérité et de « toucher » la machine comme on « touche » le piano (*voir fig. 11*). L'appareil est pourvu de cinq

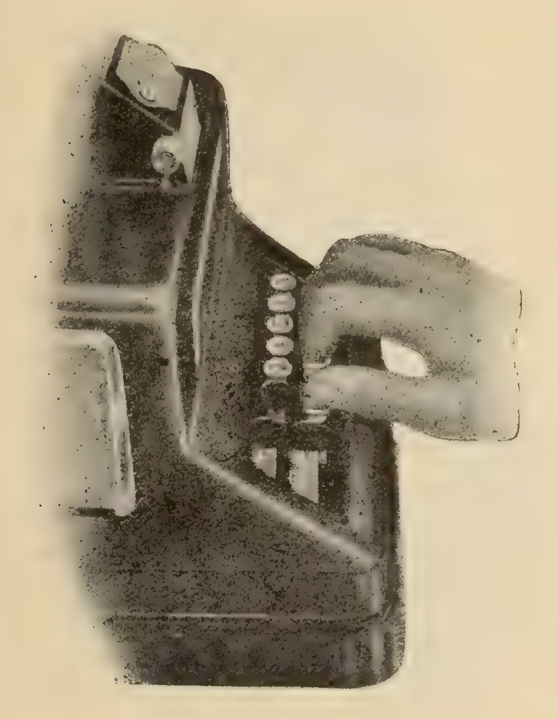


Fig. 11.

touches spéciales pour la totalisation, la répétition, la correction, l'élimination ou la signalisation des données numériques qui lui sont confiées.

Les machines à additionner que nous avons passées en revue comportent l'enregistrement écrit des données, indépendamment des résultats qui apparaissent aux lucarnes; nous avons fait ressortir l'avantage de ce dispositif pour les bureaux de statistique où il est indispensable de contrôler le travail de nombreux employés. Mais, si l'on peut

avoir en l'opérateur une confiance suffisante pour le faire travailler sans autre contrôle que celui résultant des combinaisons de colonnes, les machines dans lesquelles la manœuvre se réduit au simple enfoncement des touches présentent un avantage de rapidité évident. Telle est la machine



Fig. 12

de Felt et Tarrant, portant le nom de « comptomètre », reproduite à la *figure 12* ci-contre. Le clavier se compose de 9 rangées de chiffres comprenant chacune 8 touches. Pour additionner, il suffit d'enfoncer les touches correspondant aux chiffres des nombres à enregistrer successivement, en observant l'ordre des unités, dizaines, centaines, etc. Les résultats successifs se marquent dans les colonnes placées au bas de l'appareil.

Pour multiplier, on commence par enregistrer le multiplicande, puis on enfonce les touches autant de fois que l'exige le chiffre des unités du multiplicateur; pour multiplier par les dizaines, on répète l'enregistrement du multiplicande en reculant d'un rang vers la gauche et on enfonce les touches autant de fois qu'il le faut d'après le chiffre des dizaines du multiplicateur (1).

178. Toutes les machines à additionner se prêtent, du reste, à la multiplication par additions successives, mais il existe un avantage de rapidité en faveur des machines, comme le comptomètre, où il n'y a qu'à garder les doigts sur les touches en les enfonçant successivement un certain nombre de fois. La multiplication des nombres 6.327×6.457, par exemple, dont le produit est 40.853.439 pourrait être faite avec le comptomètre en cinq secondes. Ce sont des résultats vertigineux, tenant plus du sport que du calcul.

Les machines à additionner peuvent être aussi employées à la soustraction, grâce à l'usage des nombres complémentaires. « On appelle complément d'un nombre, dit M. d'Ocagne, celui qu'il faut lui ajouter pour atteindre la puissance entière de 10 immédiatement supérieure; ce complément s'obtient en prenant, à partir de la gauche, le complément à 9 de chacun des chiffres, sauf pour le dernier, à droite, dont on prend le complément à 10 » (2). Le complément de 3.258 est donc 6.742; pour soustraire un nombre d'un autre nombre, il faut diminuer le plus grand de la puissance 10 immédiatement supérieure et ajouter au reste le complément du chiffre le plus petit. Ainsi 48.984 — 3.258 = 38.984 + 6.742 (45.726), résultat qu'on obtient au moyen de l'addition et du calcul mental. Le comptomètre porte sur son clavier, en caractères plus petits, les chiffres com-

---

(1) On trouvera une description mécanique du comptomètre dans JACOB : *Le calcul mécanique*, Paris, 1911, pp. 30-38.

(2) Cfr. M. D'OCAGNE, *loc. cit.*, p. 37.

plémentaires jusqu'à 9, ce qui contribue à la facilité d'exécution du calcul indiqué (*fig. 12*). La répétition de la soustraction conduit à la division, comme les additions successives permettent de réaliser la multiplication. La soustraction est une opération qui intervient plutôt rarement en statistique, sauf dans le calcul des écarts. Quant à la division, elle est, au contraire, fort fréquente. Mais il est plus pratique de recourir aux machines construites spécialement pour multiplier et pour diviser : chaque machine doit être employée en vue du but pour lequel elle a été construite; ce principe s'applique aux machines à calculer comme aux machines industrielles.

179. Les machines à multiplier directement sont nombreuses. La première en date est celle inventée en 1820 par le financier Thomas, de Colmar. Cet appareil, très répandu, a été décrit en détail par M. d'Ocagne, dans son ouvrage sur « le calcul simplifié » (1). Nous ne répéterons pas ici la description du savant mathématicien français, surtout que l'arithmomètre Thomas a été remplacé dans beaucoup de bureaux de statistique par des appareils plus récents.

Plusieurs machines à multiplier sont basées sur le principe appliqué pour la première fois d'une façon pratique par le Russe Odhner. La roue à nombre variable de dents, employée par Odhner, remplace les tambours à neuf dents d'inégale longueur en usage précédemment. Les dents peuvent glisser dans des entailles convergeant vers le centre. M. Jacob en décrit comme suit le fonctionnement : « Ces dents sont munies d'une saillie, qui les oblige à rester dans une rainure circulaire formant *deux* parties de rayons différents. En faisant tourner la partie extérieure de la roue par rapport à la partie centrale, à l'aide d'un levier, les saillies s'engagent dans l'une ou l'autre partie de la rainure, ce qui fait sortir ou rentrer les dents suivant

---

(1) Cfr. M. D'OCAGNE, *loc. cit.*, pp. 45-54.

qu'elles sont en face de la grande branche ou de la petite. » Ce dispositif a l'avantage d'être peu encombrant et de créer une résistance moindre lors de la mise en marche. La roue Odhner est l'organe qui caractérise essentiellement les machines à calculer connues en France sous le nom de « Dactyle », en Allemagne sous celui de « Brunsviga », « Berolina », etc.

Pour multiplier un nombre par un autre, il suffit d'inscrire le multiplicande à l'aide de leviers glissant dans les rainures du tambour, et qu'on arrête en regard des chiffres convenables les unités à droite, puis les dizaines, les centaines, etc., vers la gauche. On donne autant de tours de manivelle de bas en haut qu'il y a d'unités à multiplier, puis on avance d'un cran la partie mobile guidée par une crémaillère, et on commence la même opération pour les dizaines. Le résultat se marque dans les lucarnes. On remet à zéro en tournant la clef placée à l'extrémité de la partie mobile. La division s'opère tout aussi facilement que la multiplication. Ces machines ont été pourvues, au cours de ces dernières années, d'une série de perfectionnements qui en rendent l'usage facile et sûr. Ainsi, une des erreurs commises le plus fréquemment, consiste à donner un tour de manivelle en trop : il suffit de faire un tour en arrière pour ramener les résultats à leur position antérieure. Si dans une opération soustractive, on retire d'un nombre un autre qui lui est supérieur, une sonnerie avertit de l'erreur commise. Enfin, les modèles nouveaux sont pourvus d'un mécanisme remettant à zéro tous les leviers à la fois. Il existe aussi des types imprimant les résultats. M. Jacob a donné une description complète de la « Brunsviga », que le lecteur curieux de détails mécaniques consultera avec intérêt (1).

La machine de M. Otto Steiger (1892), dénommée « la mil-

(1) Jacob, *loc. cit.*, pp. 63-76.

lionnaire », effectue les quatre opérations fondamentales avec une grande rapidité. Le principal avantage de cette machine consiste dans la simplicité de l'opération la plus fréquemment usitée, la multiplication. Alors que, avec les autres machines, l'opérateur doit donner autant de tours de manivelle qu'il y a d'unités de chaque ordre dans le multiplicateur, il suffit ici de placer une aiguille mobile sur le chiffre correspondant à ce nombre d'unités et de donner un seul tour de manivelle. Pour multiplier un nombre par 985, par exemple, il faut exécuter, avec les machines ordinaires, 22 tours de manivelle et, avec « la millionnaire », il n'y en a que trois (*fig. 13*).

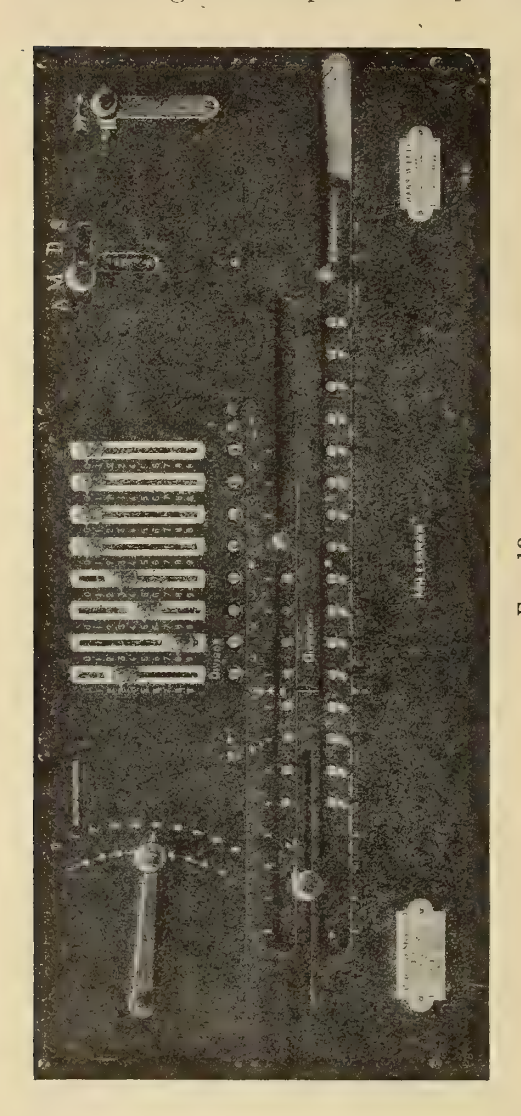


Fig. 13

Il y a un grand nombre de machines à multiplier et à diviser, dont quelques modèles fonctionnent à l'électricité; ces derniers sont doués d'une grande rapidité et sont précieux pour le calcul des pourcentages lorsqu'on en a un

grand nombre à déterminer. En poussant le calcul jusqu'à la troisième décimale, on arrive à des résultats d'une exactitude complète.

180. Ainsi que le dit M. d'Ocagne, pour avoir l'équivalent d'une règle à calcul de grande longueur, il suffit de fractionner les deux parties de la règle (la règle et la réglette) en un même nombre de parties égales et d'en placer les segments les uns en dessous des autres en les faisant alterner, après avoir eu soin de les rendre solidaires. En plan, on obtient ce qu'on appelle une grille à calcul ; en dévelop-

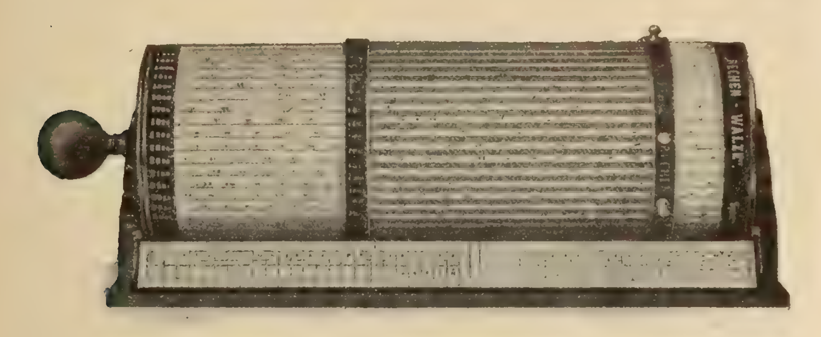


Fig 14.

pant cette grille sur la surface d'un cylindre, on obtient les cylindres à calcul.

C'est à cette dernière catégorie qu'appartient le rouleau-calculateur de M. Billeter (*fig. 14*). Cette machine à diviser, d'un usage remarquablement pratique pour les calculs de proportions, se compose d'un cylindre métallique autour duquel se meut un second cylindre nommé « curseur », composé d'un réseau de cinquante règles parallèles ayant la longueur d'une règle à calcul ordinaire. Cette disposition permet d'établir des divisions beaucoup plus nombreuses et d'obtenir des résultats beaucoup plus précis qu'avec la règle à calcul ordinaire ; tandis que cette dernière ne donne que trois chiffres, dont deux exacts et le troisième approximatif, le rouleau

nous fournit cinq chiffres dont les quatre premiers sont exacts et le cinquième est à évaluer. Le maniement, qui demande un peu d'habitude pour la lecture, est fort simple : il consiste à placer le curseur à un certain endroit du cylindre. Le fonctionnement est particulièrement rapide, quand il s'agit de diviser une certaine quantité en parts proportionnelles, cas fort fréquent en statistique : il suffit, en effet, de placer une fois le curseur à l'endroit convenable et de procéder à la lecture successive de chacun des résultats.

III. — Avantages des machines.

181. Un grand bureau de statistique ne pourrait pas plus se passer de machines à dépouiller et à calculer qu'un établissement industriel moderne de moteurs et d'un outillage mécanique. On a pu, dans certains cas, continuer de sacrifier aux vieilles coutumes, mais ce fut grâce au système de décentralisation ; quand les opérations statistiques sont centralisées, la nécessité de l'outillage moderne s'impose et emporte les dernières résistances. Celles-ci viennent de chefs trop attachés à l'organisation à laquelle ils sont doucement accoutumés ; elles viennent aussi des employés qui, souvent, ont à l'égard des machines les mêmes préjugés que les ouvriers. Ces faibles résistances ne pourront entraver le progrès technique, qui, par l'emploi des machines, s'affirme dans les directions énumérées ci-après :

1° Les machines à dépouiller et à calculer rendent possibles des combinaisons de données plus nombreuses que celles qu'il est permis de réaliser par les anciens procédés. Les travaux de statistique ont des limites déterminées par le temps et l'argent dont le bureau dispose. L'organisation qui économise la durée et la dépense augmente le domaine de la statistique et présente, en conséquence, une importance scientifique de premier ordre. M. Robert P. Porter le

constate en ces termes, à l'occasion du premier Census américain dépouillé à l'aide de la machine électrique de Hollerith, celui de 1890 : « Grâce à l'usage des machines électriques à dépouiller, il a été possible, dans le présent recensement, pour la première fois dans l'histoire du travail statistique, de retirer des bulletins toutes les informations qu'ils contenaient, dans toutes les directions désirables » (*Introduction,* Census de 1890, p. XVII, compendium).

2° Le travail mental exigé par le calcul deviendrait écrasant dans les statistiques modernes. Il ne s'agit plus ici de quelques additions, mais de centaines de milliers d'opérations arithmétiques, multiplications et divisions. Il faut disposer d'appareils se contrôlant eux-mêmes, et qui épargnent au personnel la peine de longs calculs. « Combien d'observations précieuses, a écrit un savant mathématicien, le général Menabrea, restent inutiles aux progrès des sciences, parce qu'il n'y a pas de forces suffisantes pour en calculer les résultats. » (1).

Les machines décuplent les forces dont chaque bureau dispose : elles contribuent donc directement au progrès scientifique. Les logarithmes peuvent être assimilés aux machines à calculer; Kepler fut l'un des premiers et des plus fervents adeptes de ce mode nouveau de calcul et c'est grâce aux logarithmes qu'il put dresser les tables d'où il devait faire sortir les lois des mouvements planétaires; « de sorte que, dit M. d'Ocagne, après avoir cité cet exemple, si les logarithmes n'eussent pas été inventés à propos (par Neper, en 1614), la loi de la gravitation universelle serait peut-être encore à découvrir » (2);

3° En débarrassant le chef du bureau de statistique du souci de l'exactitude matérielle des calculs effectués par ses

(1) *Notions sur la machine analytique,* de Ch. BABBAGE (Bibliothèque universelle de Genève, t. XLI, p. 352).
(2) M. D'OCAGNE, *loc. cit.,* p. 101.

collaborateurs et en lui permettant de réduire sa participation aux travaux matériels à un simple contrôle, l'emploi des machines lui assurera la faculté de se consacrer davantage à la science et de donner plus de temps à la mise en valeur des données statistiques. Ici encore, la machine concourt à un but scientifique supérieur;

4° Les résultats financiers de l'emploi des machines se modifient, sans aucun doute, à chaque expérience; trop d'éléments variables interviennent pour qu'il n'en soit pas ainsi. Mais on ne peut mettre en doute que ces résultats ne soient tous dans le sens d'une importante réduction des dépenses. Sans doute, il ne faut pas, en pratique, apprécier le rendement des machines d'après quelques éléments types fournis par les constructeurs et vendeurs de ces appareils. Sans contester la sincérité des résultats produits, on peut croire qu'il s'agit de l'effort tout momentané de quelque virtuose; c'est un record et non une moyenne de résultats obtenus après de longues semaines de travail. Mais c'est un fait d'expérience qu'une bonne machine à calculer effectue, sans erreur, le travail de plusieurs excellents employés, et c'est un résultat assez satisfaisant pour faire intervenir le travail à la machine dans la plus large mesure possible.

CHAPITRE IV

La présentation des résultats statistiques

I. — Règles pratiques de la présentation statistique.

182. La présentation statistique consiste dans le groupement et l'exposition des données en vue de la publication. Elle a pour objet de grouper les données de même nature, de telle façon que, par leur réunion, l'esprit soit frappé de certains résultats; en même temps qu'elle place les résultats sous les yeux du chercheur, elle dispose le matériel en

sorte de permettre de nouvelles combinaisons de données et de nouvelles recherches interprétatives.

C'est à l'expérience à apprendre la manière de présenter les données statistiques en tableaux clairs, bien ordonnés, ni trop simples, ni surchargés. Cet art ne peut guère s'enseigner, mais il est un ensemble de recommandations dont on fera bien de ne pas s'écarter si l'on tient à rester dans la norme des publications ayant une bonne tenue scientifique.

A quel moment faut-il s'occuper de la préparation des cadres de publication? Nous répondons, sans hésitation : dès le début du travail, avant même de rédiger le questionnaire ou bulletin. Il faut savoir où l'on va pour ne pas surcharger le bulletin de questions inutiles, ni omettre l'un ou l'autre point essentiel. Dresser un schéma du cadre de publication est la meilleure préparation qu'on puisse imaginer à la rédaction du bulletin. Puis, après la critique, on verra quelles sont les données utilisables et on décidera dans quelle forme elles doivent être publiées : c'est la phase définitive de la préparation des cadres. Les différentes parties de la statistique sont ainsi solidaires : le contenu du bulletin dépend du but de la statistique, qui est exprimé dans les cadres et les cadres eux-mêmes dépendent des réponses du bulletin, modifiées par la critique. Les trois parties de la statistique que nous avons exposées jusqu'ici se compénètrent donc réciproquement.

Le principe essentiel en la matière est l'ordre et la clarté. On ne perdra pas de vue que les statistiques sont faites pour fournir des renseignements au public et non pour exercer l'ingéniosité du lecteur en lui posant des énigmes. Tout, dans la présentation, doit obéir à ce principe fondamental : énoncé des titres, disposition des colonnes, rédaction des notes. Un tableau bien fait doit se lire facilement, les recherches doivent y être aisées, on doit ressentir, en le regardant, une impression d'ordre, comprendre qu'on se trouve devant un plan logiquement ordonné. Nous passe-

rons rapidement en revue quelques points se rattachant à cette question générale.

A) Faut-il un cadre unique, ou est-il préférable d'en composer plusieurs? Il est rare qu'une statistique, tant soit peu complexe, puisse tenir en un cadre unique. Si l'on veut s'astreindre à cette règle, on court risque d'aboutir à une grande confusion, de manquer absolument de clarté. Un cadre doit avoir son objet propre et présenter au complet les données numériques concernant cette question. En général, il est donc préférable qu'une statistique soit présentée en plusieurs cadres, chacun d'eux étant relatif à une question ou à un groupe de questions.

B) Chaque tableau doit avoir son titre et ce titre doit indiquer exactement, sans omission et sans ajoute, le contenu du tableau. Le lecteur doit être informé, en lisant le titre seul, de ce qu'il va trouver dans le cadre statistique. Non seulement, l'objet matériel doit être indiqué, mais encore le point de vue sous lequel il est envisagé. Si l'on éprouve quelque difficulté à libeller cet énoncé, ce peut être un indice que le cadre est surchargé. Ces remarques s'appliquent aussi aux intitulés des colonnes.

C) En général, une donnée statistique pour acquérir toute sa signification, doit être présentée en nombres absolus et en pourcentages. Mais faut-il réunir ces deux espèces de données en un cadre unique, ou consacrer un cadre spécial à chacune d'elles? Dans les tableaux statistiques qui forment l'exposé systématique du matériel, les tableaux de publication proprement dits, il ne faut pas de chiffres proportionnels. On se borne là à montrer les résultats tels qu'ils ont été calculés, on n'a pas à les comparer à d'autres. Il en va autrement dans les tableaux d'analyse : là, les chiffres proportionnels trouvent un emploi constant et sont un moyen d'investigation indispensable. Nous ne sommes pas partisan de les publier isolément, car un pourcentage séparé du nombre absolu qu'il représente est souvent trompeur.

Une augmentation relative peut être énorme et ne correspondre qu'à un chiffre absolu de fort médiocre importance. Le pourcentage isolé nous induit en erreur plutôt qu'il ne nous éclaire.

D) Le nombre de colonnes dépend d'un fait matériel ; le format de la publication, et de deux autres éléments : la complexité de la matière et le détail dans lequel on désire entrer. On peut adopter, pour règle générale, de subdiviser la question en partant de la classe la plus générale pour arriver à la plus spéciale, comme dans l'exemple ci-après, où il s'agit de la répartition d'une population selon l'âge, le sexe et l'état civil, dans les villes et les campagnes.

1.	Province de Brabant											
2.	Localités urbaines.						Localités rurales.					
4.	Hommes.			Femmes.			Hommes.			Femmes.		
16.	Cél.	mar.	veufs div.	Cél.	mar.	veuves div.	Cél.	mar.	veufs div.	Cél.	mar.	veuves div.

18 ans.
19-30 »
30-50 »
50-60 »
60 et plus.

La présentation par tableau à double entrée, comme l'exemple ci-contre, équivaut à multiplier le nombre des colonnes : le résultat est le même que si chaque colonne était partagée en cinq subdivisions, ce qui ferait 5 × 16 ou 80 colonnes.

E) Une autre question, fort intéressante, est celle-ci : quelles sont les données qui devraient être publiées isolément ? Faut-il, dans la statistique d'un pays, descendre jusqu'à l'unité géographique la plus faible, la commune, ou peut-on s'arrêter en deçà ? C'est un problème qui se pose à tout instant et qui a soulevé de vives controverses. Nous pensons que sa solution n'a rien d'impossible. En principe, il faut descendre dans les plus petits détails, parce qu'en dehors de ce qui est publié, le chercheur ne peut pas disposer du matériel rassemblé par le statisticien. Mais, il ne faut pas que ce souci aille jusqu'à la minutie. On peut, on doit adopter

des règles dont l'effet est de réduire la longueur, partant le coût et la durée des publications. Ainsi, pour le recensement industriel belge de 1910, nous avions pris pour règle de ne citer que des communes où, dans l'industrie considérée, dix personnes au moins travaillaient (patrons et ouvriers réunis). Les communes non citées étaient réunies sous la dénomination « autres communes ». Ce simple principe d'élimination, qui n'enlève rien de leur valeur aux données statistiques, a eu pour effet d'alléger la publication de plusieurs milliers de lignes dénuées d'intérêt.

183. La forme matérielle de la présentation peut aussi faire l'objet de quelques remarques.

A) Le format de la publication doit être l'objet d'une sérieuse attention. Les formats trop petits ne se prêtent pas à l'impression de tableaux statistiques, à moins d'employer des caractères microscopiques, fatigants pour tout le monde, illisibles pour quelques-uns. On adopte généralement l'in-8° pour les « Annuaires », qui publient surtout des tableaux résumés, et on réserve l'in-4° aux grandes publications statistiques. Les formats plus grands se placent difficilement dans les bibliothèques : sauf des cas spéciaux, comme lorsqu'il s'agit de collections commencées depuis longtemps, les grands formats, comme les in-folio, ne sont pas à recommander.

B) Un tableau statistique peut être imprimé sur deux pages à condition que le repérage des lignes soit très exact, mais jamais sur quatre. Pour pouvoir consulter utilement une statistique, il faut que l'on ait sous les yeux, à la fois, toutes les données du tableau.

C) Le papier employé doit être mince et résistant. Il ne faut pas viser à l'effet en grossissant les volumes ; il faut songer que les publications volumineuses font l'effroi des bibliothécaires.

D) Un tableau statistique doit être rempli, c'est-à-dire

que toutes ses colonnes doivent être fournies de chiffres. S'il en est autrement, revoyez attentivement votre classification de base et vos en-têtes de colonnes : il y a quelque part des divisions plus minutieuses qu'il ne conviendrait.

E) Lorsque la page est couverte de chiffres, on obtient un bon résultat pour la clarté en laissant une ligne de blanc de cinq en cinq, ou de dix en dix lignes. Lorsque les résultats sont présentés par années, on peut introduire cette ligne de blanc après chaque décade.

Dans un but identique, on peut numéroter les données et répéter ces indications à droite et à gauche du cadre. Le nom des localités ou les chiffres des années peuvent aussi faire l'objet d'une double inscription — à gauche et à droite.

F) Les totaux dans les publications anciennes sont généralement placés à la fin des colonnes et le total général qui résume les totaux partiels des colonnes se trouve à droite. Partant de cette idée que le total est le chiffre qu'on cherche en premier lieu, les statisticiens anglais et américains ont pris pour règle de placer la colonne des totaux à gauche, et même de faire précéder les chiffres de détail, du total général pour le Royaume. L'innovation peut avoir du bon, mais il ne peut être question de condamner le procédé ancien.

Si les données doivent être comparées, il est recommandable de rapprocher les colonnes de totaux se rapportant à ces données.

G) Les colonnes doivent être combinées de façon à assurer une série de contrôles permettant de vérifier l'exactitude matérielle du dépouillement et des additions. Lorsque ces contrôles sont bien établis et qu'on prend la peine de les vérifier avec attention, les erreurs sont infiniment réduites. Les employés, sachant que leur travail est contrôlé, y mettent d'autant plus de soin.

H) Les « coquilles » sont toujours désagréables, mais dans un travail de statistique elles sont désastreuses. On ne

peut donc apporter trop de soins à la correction des épreuves, ni montrer trop d'exigences à l'égard de la capacité professionnelle de l'imprimeur. Les « bons à tirer » doivent être revus avec attention par une personne qualifiée, le chef du bureau lui-même, s'il se peut.

184. — *Références.*

« American Census taking » (*Century Magazine*), 1903.

BERTILLON (J.), *Cours élémentaire de statistique administrative*, Paris, 1895, pp. 62-72 et Appendice.

BERTILLON (J.), « La statistique à la machine » (*La Nature*, 1er septembre 1894, n° 1109). ID., *ibid.*, janvier 1914.

BLOCK (M.), *Traité théorique et pratique de statistique*, deuxième édition, Paris, 1886, pp. 292-306.

BOSCO (A.), *Lezioni di statistica.* Parte prima : « Metodologia statistica », pp. 256-290.

BOWLEY (A. L.), *Elements of statistics*, second edition, London 1902, pp. 73-103.

BOWLEY (A.L.), *An elementary manual of statistics*, London, 1910, pp. 50-55.

CHEYSSON (E.), « La machine électrique à recensement » (*Journal Soc. Stat.*, Paris, 1892, p. 87).

D'OCAGNE (M.), *Le calcul simplifié par les procédés mécaniques et graphiques*, Paris, 1905.

DURAND (E. B.), « Census methods » (*American Statistical Association*, 1908-1909, pp. 608 et suivantes).

GABAGLIO, *Teoria generale della statistica*, Milano, 1888, t. II, pp. 410-12.

HOLLERITH (H.), « The electrical tabulating machine » (*Journal of the Royal Statist. Society*, London, 1894, p. 678).

HOOKER (R. H.), « Modes in census taking in the British Dominions » (*Journal of the Royal Statist. Society*, London, 1894, pp. 289 et suiv.)

JACOB (L.), *Le calcul mécanique*, Paris, 1911.

JULIN (A.), *Précis du cours de statistique générale et appliquée*, quatrième édition, Bruxelles et Paris, 1919, pp. 58-64.

KING (W. J.), *The elements of statistical method*, New York, 1912, pp. 83-90.

NEWCOMB (H. T.), *Mechanical tabulation of the statistics of agriculture on the twelfth Census of the U. S.*, Philadelphia, 1901.

PIDGIN (C. F.), *Practical statistics*, Boston, 1888, pp. 147-160.

RAUCHBERG (H.), « La machine électrique à recensement. Expériences et améliorations » (*Bulletin de l'Institut international de statistique*, t. IX, première livraison, Rome, 1895, p. 249).

VITOUX (G.), « Le classi-compteur-imprimeur » (*La Nature*, 11 mai 1901, n° 1459).

VON MAYR (G.), *Statistik und Gesellschaftslehre*, Tubingen, 1914. pp. 106-134.

YULE (Udny G.), *An Introduction to the theory of statistics*, London, 1911, pp. 11-15.

LIVRE SECOND

PROCÉDÉS D'ANALYSE DU MATÉRIEL STATISTIQUE

Généralités et division de la matière

185. Au cours de l'exposé compris dans les trois sections qui composent le livre premier, nous avons passé en revue les phases du travail statistique en tant qu'elles concernent le relevé, la critique, le dépouillement et la présentation des données. On pourrait appeler cette partie la « formation des tableaux primaires ». Le dépouillement et la présentation ne font que placer sous les yeux du lecteur les résultats bruts de l'observation, en les classant d'après les divisions des nomenclatures et en les totalisant par groupe ou classe. Mais une phase nouvelle va s'ouvrir. Il ne suffit pas d'exposer, après l'avoir classé, le matériel statistique. Cette suite, déjà longue, d'opérations que nous avons décrite, pourrait se comparer au travail du naturaliste qui recherche des types d'animaux, s'efforce de les saisir dans leurs conditions de vie, de les présenter *in situ* et puis les range, par espèces et familles, dans les vitrines d'un musée. L'histoire naturelle n'eût fait que des progrès bien lents si l'on s'en était tenu à cette science purement descriptive. Aussi la

tâche des conservateurs et du directeur d'un musée d'histoire naturelle ne s'arrête pas là. Il leur reste à mettre en œuvre les matériaux recueillis, à les comparer, à rechercher les relations des êtres entre eux (rapports statistiques, dynamiques et organiques), à étudier la variation des êtres dans leur nature, leur origine et leur fin, à tirer les conclusions d'ordre scientifique qui se dégagent de cet examen (1). La seconde partie des opérations statistiques que nous examinons au cours du présent livre ressemble à ce travail d'analyse. Le statisticien, comme le naturaliste, doit aussi comparer entre eux les matériaux mis à sa disposition — pour lui, ce sont les données statistiques formées à l'aide des calculs du dépouillement — et il faut trouver les expressions synthétiques qui résument, le plus exactement possible, la masse des constatations numériques. Le langage que tiennent les chiffres est plein d'obscurités et d'embûches : l'égyptologue, courbé sur ses palimpsestes, n'a guère plus de difficultés à vaincre que le statisticien n'en rencontre au cours de l'analyse de son matériel. Ce n'est que par là cependant qu'il arrive à dégager les traits généraux, à marquer les ressemblances, à mesurer les différences, à présenter à l'homme d'étude les éléments simples dont celui-ci a besoin pour rechercher les tendances régulières et les causes des phénomènes.

L'exposé de ces procédés d'investigation scientifique forme la matière de notre second livre.

186. On pourrait la diviser en tenant compte de la nature des opérations mathématiques à effectuer et considérer d'abord les procédés d'analyse basés sur l'arithmétique, puis ceux fondés sur l'algèbre et enfin ceux qui se démontrent par la géométrie. Dans la première catégorie

---

(1) Voyez, par exemple, le tableau synoptique résumant le programme de travail en histoire naturelle, inséré à la fin du mémoire de G. GILSON, déjà cité (Bruxelles, 1914).

rentrent les moyennes, les proportions et rapports, etc.; à la seconde appartiennent l'interpolation et la théorie de la corrélation; dans la troisième se rangent les procédés divers de la statistique graphique.

Mais il serait permis de reprocher à cette classification quelque chose d'arbitraire, car les formules que nous rencontrerons au cours de notre exposé peuvent en général se démontrer à la fois par l'algèbre et la géométrie.

Nous avons préféré conserver, dans ce livre, comme nous l'avions fait dans le livre premier, l'ordre logique des opérations statistiques, car nous écrivons pour les statisticiens, avant tout.

D'après cet ordre, nous considérerons d'abord les séries et leurs modalités, ainsi que les différentes formes de distribution. Puis nous envisagerons les mesures diverses des séries : les moyennes et le calcul de la dispersion. Nous étudierons ensuite la manière de fixer le rapport de dépendance des phénomènes; c'est la théorie des corrélations. Nous verrons, après cela, ce qui a trait à la statistique graphique. Nous terminerons enfin par l'exposé de la loi des erreurs.

Il n'est pas inutile de rappeler ici le point de vue général que nous avons exposé dans notre introduction sous le titre de : « La statistique et les mathématiques ».

Sans aucun doute, les mathématiques et la statistique ne se confondent pas et on a vu de bons statisticiens sans connaissances plus approfondies que les mathématiques élémentaires. Mais les domaines de ces méthodes scientifiques se touchent et leurs fonctions sont communes, si bien qu'on ne peut décrire toute la méthode statistique sans en venir à toucher le domaine des mathématiques. Ceci est particulièrement vrai à propos de la matière traitée au cours des livres II et III. Si l'on peut faire en quelque sorte abstraction des mathématiques quand il s'agit du but concret de recueillir, ordonner, dépouiller et publier les documents

statistiques, il n'en est plus de même lorsqu'il faut analyser, comparer et vérifier les résultats qui s'en dégagent. Encore une fois, il ne peut être question que de domaines voisins, mais ne formant pas un territoire unique. Il y a juxtaposition, non compénétration. Le statisticien a recours aux procédés mathématiques ; il s'appuye sur les théorèmes de la géométrie et de l'algèbre, il doit montrer quelle est leur signification par rapport aux phénomènes décrits par la méthode, il ne doit pas nécessairement en refaire, après d'autres, la démonstration. Cette tâche incombe proprement aux mathématiciens. Chercher et démontrer, par les procédés mathématiques, que tel résultat s'obtient par telle formule, nous semble rentrer dans le rôle du mathématicien. Faire emploi de cette formule, montrer comment elle s'applique à des phénomènes statistiques, nous paraît la tâche du statisticien. Toutefois cette distinction n'a rien d'absolu ; il est bien difficile de faire l'une ou l'autre application raisonnée sans y joindre au moins une démonstration élémentaire. On a fait remarquer avec raison que, privée de toute démonstration mathématique, la description des procédés statistiques ne serait plus qu'un recueil de « recettes », vide de toute pensée scientifique. Il faut donc s'efforcer de rester dans un juste milieu. C'est de cette pensée que nous nous sommes inspiré au cours des pages suivantes.

CHAPITRE PREMIER.

Séries, sériation, distribution

I. — Les séries statistiques.

187. Pour acquérir toute la valeur dont elle est susceptible, l'observation statistique doit être la plus longue et la plus étendue possible. La plus longue possible, afin que les faits relevés acquièrent toute leur signification en dépouillant les caractères accidentels qu'ils peuvent présenter dans certains cas isolés; la plus étendue possible, afin que les caractères généraux se réflètent dans l'observation et non pas ceux qui ne sont qu'exceptionnels.

Une succession de données résultant de l'observation quantitative d'un fait non typique, répétée dans le temps ou dans l'espace, prend le nom de série.

Les statisticiens n'ont pas toujours paru attacher une importance suffisante à la question de la classification des séries. Cependant, si l'on veut y réfléchir un instant, on s'apercevra que les résultats de la statistique d'investigation sont très différents selon qu'ils s'appliquent à une espèce de séries, ou à une autre espèce.

Ou bien les phénomènes sont observés par la statistique d'après des divisions du temps, ou bien ils le sont selon des divisions de l'espace.

Ceci fournit une base de classification rationnelle. Nous aurons en premier lieu des séries dans lesquelles l'évolution du phénomène est considérée dans le temps; ce seront les séries chronologiques ou historiques.

D'autre part, nous distinguerons les séries dans lesquelles l'élément de succession dans le temps fait défaut, mais dans lesquelles l'observation statistique, supposée

faite à un moment donné, s'applique à des endroits différents du territoire. Ces séries montrent les variations du phénomène dans l'espace.

Mais certains phénomènes peuvent être étudiés en eux-mêmes, sans considération du temps et de l'espace, mais simplement dans le but de montrer la distribution des éléments variables dont ils se composent. Et ceci nous fournit une nouvelle base de classification des séries.

Somme toute, la classification proposée n'a rien de nouveau dans la statistique. Nous la retrouverons dans l'exposé de la statistique graphique. Bien qu'il soit peu désirable, dans un exposé méthodique, d'anticiper sur les développements de la matière, nous devons signaler ici que la statistique graphique divise les diagrammes en deux classes : les diagrammes historiques ou chronologiques qui représentent les faits d'après leur évolution au cours des années, et les diagrammes de distribution qui montrent la répartition des fréquences, sans intervention de la notion de temps. A ces deux modes de réprésentation, la statistique graphique en ajoute un troisième qui envisage la répartition dans l'espace : les cartogrammes.

Les divisions sont exactement celles proposées à l'égard des séries (1).

188. Le premier groupe est formé des séries basées sur une mesure du temps.

L'unité de temps admise est souvent l'année. On peut adopter l'année astronomique ou l'année fiscale ou financière. Cette dernière unité n'est employée que dans des statistiques spéciales ayant trait aux finances publiques ou privées. Il y a toujours lieu de préciser la nature de l'unité de temps dont on se sert. Les

---

(1) Cette classification est celle proposée par M. Edmond F. Day dans un article paru dans « Quarterly publications of the American Statistical Association », vol. XVI, n° 128 (1919) ; nous l'avions nous-même adoptée en partie dès 1917.

statistiques ayant l'année comme mesure sont extrêmement nombreuses. Parmi elles, on peut citer, à titre d'exemple, les statistiques criminelles; nous en donnons ci-après un spécimen :

EXEMPLE 1.

ANNÉES	NOMBRE DE CONDAMNÉS (1)						
	HOMMES			FEMMES			TOTAL des hommes et des femmes
	Primaires	Récidivistes	TOTAL	Primaires	Récidivistes	TOTAL	
1900	23,099	18,134	41,233	9,203	3,251	12,454	53,687
1905	20,024	19,334	39,358	8,548	4,141	12,689	52,047
1906	21,307	20,410	41,717	8,432	4,448	12,880	54,597
1907	20,446	20,522	40,968	8,120	4,417	12,537	53,505
1908	19,919	20,016	39,935	8,184	4,208	12,392	52,327
1909	18,674	18,915	37,589	8,156	4,449	12,605	50,194
1910	20,443	20,269	40,712	8,135	4,573	12,708	53,420
1911	19,278	19,129	38,407	7,879	4,351	12,230	50,637
1912	20,818	20,414	41,232	8,407	4,841	13,248	54,480

L'année astronomique ou encore l'année budgétaire constitue souvent une mesure trop large pour apprécier, avec la précision voulue, certains phénomènes qui, au cours d'une même année, subissent des fluctuations périodiques, ou dont on désire suivre le développement à intervalle plus rapproché. Le commerce a intérêt, par exemple, à connaître l'allure des échanges internationaux dans le plus bref délai possible; aussi les statistiques commerciales, qui ne sont définitivement closes qu'après l'année écoulée, paraissent-elles, sous une forme provisoire et abrégée, de mois en mois. Quelques statistiques sont publiées, sous cette même forme provisoire, tous les six mois; il en est ainsi du relevé de la production et du personnel dans les mines de houille et les usines métallurgiques en Belgique.

(1) D'après la « Statistique judiciaire de la Belgique », publiée par le Ministère de la Justice. Bruxelles, in-folio (paraît annuellement).

Dans certains cas, la publication mensuelle et la formation de séries établies pour chaque mois sont d'une impérieuse nécessité, afin de suivre les variations constantes d'un fait dépendant essentiellement de l'époque à laquelle il se produit : telles sont les statistiques météorologiques. Le relevé des heures d'illumination solaire fournit un exemple typique de ce genre de séries.

EXEMPLE 2. — **Heures de soleil, à Uccle (Totaux mensuels).**

Institut météorologique de Belgique (1).

Années	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre	Total
1898	57	54	87	205	146	176	202	278	224	108	104	67	1708
1899	92	134	168	133	194	262	271	275	172	195	76	62	2034
1900	33	70	101	170	211	192	296	205	197	147	44	48	1714
1901	98	65	78	193	283	236	254	223	160	126	95	30	1841
1902	27	99	118	145	175	212	235	155	142	77	95	59	1539
1903	75	60	150	107	237	186	179	178	162	95	51	61	1541
1904	60	71	66	162	183	215	309	275	161	114	53	37	1706
1905	81	59	117	120	228	214	242	200	115	66	45	42	1529
1906	69	48	133	219	177	194	215	229	187	134	51	48	1704
1907	51	64	184	134	244	162	177	205	188	92	72	49	1622
1908	96	44	77	139	162	212	211	201	182	207	109	39	1679
1909	93	116	70	235	329	141	134	228	130	106	89	36	1707
1910	51	68	179	164	188	169	127	186	132	115	55	27	1461
1911	55	64	107	184	230	184	327	280	229	131	67	25	1883
1912	33	64	89	232	180	176	197	108	109	144	27	49	1408

189. Quelquefois, les exigences de la recherche scientifique rendent nécessaire une division du temps plus courte que le mois : la journée, par exemple. La statistique des

(1) Extrait de l'*Annuaire météorologique pour 1914*, Bruxelles, 1915.

accidents du travail a montré que les sinistres étaient plus fréquents certains jours que d'autres. Si les ouvriers n'ont pas la sagesse de vivre sobrement le dimanche, l'ébranlement nerveux produit par l'ivresse manifestera ses effets pendant la journée du lundi et provoquera ce jour-là un plus grand nombre d'accidents; il se peut aussi que la fatigue accumulée au cours de la semaine amène l'épuisement et que cet état ait pour conséquence un plus grand nombre d'accidents vers la fin de la semaine. Ces constatations ne sont possibles que par l'enregistrement des accidents, jour par jour. Bien entendu, ces chiffres, pour revêtir leur signification exacte, doivent être mis en rapport avec le nombre d'ouvriers occupés chacun des jours de la semaine, et seulement pour des journées de durée normale.

EXEMPLE 3. — **Relevé des accidents du travail par jour.**

(Allemagne, 1907) (1) :

Lundi	13,707	ou	16.44 %		Vendredi	13,173	ou	16.28 %
Mardi	12,752	ou	15.76 %		Samedi	13,603	ou	16.80 %
Mercredi	12,668	ou	15.66 %		Dimanche	2,036	ou	2.52 %
Jeudi	12,997	ou	16.4 %		Ensemble	80,916	ou	100.00 %

190. Dans certaines recherches, on est allé plus profond encore en présentant les séries statistiques d'après une unité extrêmement réduite : l'heure. On a appliqué cette échelle à la statistique des décès et des naissances. M. Raseri a pu établir que le nombre moyen des naissances et le nombre moyen des décès varient entre des limites très étendues suivant les heures de la journée. Par 1,000 décès qui se vérifient dans les 24 heures, la moyenne horaire est 49 entre 2 heures de l'après-midi et 7 heures du soir, et

(1) Données extraites de *Gewerbe-Unfallstatistik für das Jahr 1907, Erster Teil*. Berlin, 1910, tabelle II.

elle est seulement 34 entre 7 heures du matin et 2 heures de l'après-midi. Par 1,000 naissances, la moyenne horaire est 52 entre 1 heure et 8 heures du matin, tandis qu'elle est seulement de 39 entre 1 heure de l'après-midi et 8 heures du soir. Voici la répartition des décès, d'après les heures, enregistrés à Crémone de 1866 à 1880 (1) :

Exemple 4.

Heures	Décès	Proportion à 1,000	Heures	Décès	Proportion à 1,000
1	1,069	42	13	1,062	42
2	957	38	14	1,249	49
3	1,136	45	15	1,379	54
4	1,032	41	16	1,292	51
5	1,085	43	17	1,267	50
6	1,103	43	18	1,064	42
7	1,029	40	19	950	37
8	968	38	20	935	37
9	1,047	41	21	880	35
10	1,077	42	22	1,047	41
11	1,000	39	23	549	22
12	1,081	42	24	1,216	48
				25,474	1,000

La division du temps d'après laquelle la série est construite s'appelle l'échelle de la série. Cette façon de présenter les résultats statistiques est simple et commode. A la condition d'adopter une division appropriée à la recherche scientifique qu'on se propose, ce système présente le grand

(1) Enrico Raseri, « Les naissances et les décès suivant les heures de la journée » (*Bulletin de l'Institut international de statistique*, t. XI, livre I, p. 144).

avantage de n'être pas arbitraire, qualité qui ne se retrouve pas toujours dans le second mode de présentation.

191. Dans le second groupe nous rangeons les séries dans lesquelles les unités sont réparties d'après leur situation dans l'espace. La présentation par unité géographique offre naturellement moins de précision que celle basée sur les divisions du temps, parce que le choix du statisticien peut se porter à peu près indifféremment sur telle unité ou sur telle autre. L'ordre du classement n'échappe pas non plus à tout arbitraire. Il peut être basé sur la succession naturelle des nombres ou sur leur importance proportionnelle dans certaines divisions territoriales; une étude attentive de ces séries est la première condition à réaliser pour la confection des cartogrammes. Les comparaisons statistiques portant sur des époques différentes doivent être, en outre, surveillées de près; il arrive fréquemment que la composition des unités territoriales se soit modifiée dans l'intervalle : une commune urbaine a, par exemple, annexé à son territoire des communes rurales proches, etc.

A côté de cet inconvénient, les séries disposées d'après les divisions de l'espace présentent un avantage : c'est de se prêter à de nombreuses combinaisons. On peut énumérer les provinces d'un pays en suivant l'ordre alphabétique, ou en les classant d'après leur situation géographique (provinces du nord, du sud, de l'est, etc.), ou suivant l'importance de leur population, ou encore d'après tel caractère économique (provinces industrielles, agricoles). En recourant simultanément à plusieurs de ces critères, on multiplie et on précise en même temps les aspects sous lesquels les phénomènes peuvent être envisagés. Même dans les cas les plus simples, cette présentation double ou triple présente des avantages. Le rapport, pour mille, de la population industrielle dans chaque province belge à la population totale a été établi de la sorte par le recensement de l'industrie et du commerce (1910) :

EXEMPLE 5. — **Proportion (pour mille) de la population industrielle à la population totale dans les provinces ci-dessous :**

Les deux sexes réunis		Les hommes		Les femmes	
Hainaut	286	Hainaut	492	Flandre Orientale . .	174
Liége	279	Liége	466	Flandre Occidentale .	172
Flandre Orientale . .	237	Anvers	351	Brabant	99
Flandre Occidentale . .	229	Namur	347	Liége	94
Brabant	213	Brabant	333	Hainaut	75
Anvers	210	Flandre Orientale .	301	Anvers	71
Namur	196	Flandre Occidentale	288	Namur	48
Luxembourg	121	Luxembourg . . .	205	Luxembourg . . .	33
Limbourg	102	Limbourg	170	Limbourg	33

192. Le troisième groupe est formé par les séries qui ne sont basées ni sur la succession chronologique, ni sur la répartition dans l'espace. Au lieu de considérer la répétition d'un phénomène dans le temps, ou les aspects divers de ses manifestations à des endroits différents, nous pouvons étudier les faits en eux-mêmes, à un moment donné. Les unités qui composent ces séries sont de grandeurs diverses, mais elles présentent toutes un caractère spécial permettant de les grouper et de les considérer dans leur ensemble. Le salaire d'un manœuvre est quelque chose de différent du salaire de l'ouvrier qualifié; mais, dans une même profession, il y a une masse de salaires bas, moyens, élevés, dont l'ensemble caractérise *le salaire* de la profession. On voit que cette troisième espèce de séries est très différente des deux premières. Elle embrasse des cas nombreux : qu'il s'agisse de la répartition des revenus, de la grandeur des hommes adultes, ou de la distribution des âges parmi une population, c'est dans le troisième groupe que rentrent les séries constituées à l'aide de ces nombres. D'ordinaire, les divisions de ces séries dépendent de la nature des unités elles-mêmes; il peut arriver aussi que les divisions soient

déterminées par la précision avec laquelle les observations des cas individuels ont été faites. Les statisticiens anglais donnent le nom de « discrete series » aux séries citées en premier lieu, tandis qu'ils appellent « continuous series » les secondes.

La distribution des salaires parmi une population ouvrière fournira un bon exemple des séries que nous envisageons ici :

Exemple 6.

Salaire d'une journée de travail pour les hommes de plus de 16 ans appartenant aux industries des métaux.

(Salaires et durée du travail dans les industries des métaux au 31 octobre 1903). Publication de l'Office du Travail de Belgique, Analyse, p. 9.

Taux du salaire	Nombre d'ouvriers		Taux du salaire	Nombre d'ouvriers	
Francs	Nombre absolu	°/o	Francs	Nombre absolu	°/o
Moins de 1.50. .	978	1.16	De 4.25 à 4 49 .	4,545	5.40
De 1.50 à 1.74. .	1,495	1.78	De 4.50 à 4.74 .	5,413	6.43
De 1.75 à 1.99. .	1,442	1.71	De 4.75 à 4.99 .	3,031	3.60
De 2.00 à 2 24. .	2,699	3.21	De 5.00 à 5.24 .	4,153	4.94
De 2.25 à 2 49. .	2,263	2.69	De 5.25 à 5.49 .	1,706	2.03
De 2.50 à 2.74. .	4,407	5.24	De 5.50 à 5.74 .	2,081	2.47
De 2.75 à 2.99. .	4,926	5.85	De 5 75 à 5 99 .	1,036	1.23
De 3.00 à 3.24. .	9,048	10.75	De 6.00 à 6.24 .	1,367	1.63
De 3 25 à 3.49. .	7,079	8.42	De 6.25 à 6.49 .	571	0.68
De 3.50 à 3.74. .	8,657	10.29	De 6.50 à 6.74 .	644	0 77
De 3.75 à 3 99. .	6,575	7.82	De 6.75 à 6.99 .	323	0.38
De 4.00 à 4.24. .	8,071	9.59	7.00 et plus . .	1,626	1.93
			Total. . .	84,136	100.00

193. Plusieurs statisticiens ont proposé un mode de classification différent, qui serait emprunté à l'allure des nombres inclus dans la série, selon la tendance stationnaire, ascendante ou descendante qu'ils manifestent. Sans doute, aucun phénomène collectif n'est absolument stable; ses unités, essentiellement variables, impriment à la masse une agitation incessante, mais ces changements ont une direction marquée qui se peut caractériser.

Si les nombres compris dans une série présentent cette marque d'une stabilité relative, on dira de cette série qu'elle revêt un caractère constant; peu importe que les écarts entre les données soient presque insignifiants, ou que, après quelque brusque déviation, les données reviennent au niveau antérieur, comme orientées vers une moyenne.

La répartition proportionnelle, suivant l'âge et suivant l'état de famille, aux divers recensements, fournit un exemple de séries constantes; la régularité est plus grande en ce qui concerne la répartition suivant l'âge, mais celle que présente la répartition selon l'état civil est encore suffisante pour la faire rentrer dans la catégorie des séries stables. Pour ne pas alourdir le tableau, nous ne donnons les chiffres que pour certaines catégories d'âges :

Exemple 7.

France (1). Répartition proportionnelle suivant l'âge et l'état civil aux divers recensements.

Années des recensements	Série constante statique — Population présente de 40 à 59 ans		Série constante statique — Population présente mariée		Série constante irrégulière — Veufs-ves, divorcés-cées	
			de 60 ans et plus	de 50 ans et plus	de 18 à 59 ans	de 15 à 49 ans
	Hommes P. c.	Femmes P. c.	Hommes P. c.	Femmes P. c.	Hommes P. c.	Femmes P. c.
1851	22.59	22 69	64.71	53 55	3.60	4 07
1856	23.12	22 58	64 15	53 88	3.99	4.61
1861	22.87	22.58	64.87	54.22	3 75	4 34
1866	22.90	22 78	64 56	54.25	3.78	4.47
1872	23.04	22.89	63.09	53 81	4 26	5.72
1876	22 68	22.79	63 96	53.12	3 95	5.43
1881	22.58	22 74	63.48	53.12	3 97	4 83
1886	22.27	22 25	63.78	53 45	4.01	5 08
1891	22 39	22 45	64.22	52.39	4.16	5 11
1896	22.31	22.34	64.33	51 32	4 03	5.08
1901	22.44	22.65	64 40	48.42	3.73	5 21

(1) Données extraites de l'*Annuaire statistique de la France*, 1911, partie rétrospective, pp. 12-13.

194. L'appellation de « série dynamique régulière » est réservée aux séries où se manifeste une tendance à la modification d'un état de choses antérieur, tendance constante agissant avec une régularité plus ou moins complète, mais d'une façon continue. Ces séries peuvent se composer de nombres disposés en ordre croissant et alors on les appelle séries dynamiques à ordre croissant, ou l'inverse se produit et dans ce cas on leur donne le nom de séries dynamiques décroissantes. Un grand nombre de phénomènes économiques se marquent sous la forme de séries dynamiques croissantes.

EXEMPLE 8.

Série dynamique croissante		Série dynamique décroissante		Série dynamique croissante		Série dynamique dcéroissante	
ANNÉES	BELGIQUE Moteurs à vapeur, force en HP.	ANNÉES	BELGIQUE Proportion des miliciens ne sachant ni lire, ni écrire.	ANNÉES	BELGIQUE Moteurs à vapeur, force en HP.	ANNÉES	BELGIQUE Proportion des miliciens ne sachant ni lire, ni écrire.
1885	781,755	1876	18.41	1897	1,208,479	1888	13.29
1886	793,924	1877	19.97	1898	1,249,813	1889	13 13
1887	812,980	878	18.62	1899	1,312,319	1890	13.05
1888	821,988	1879	18 81	1900	1,388,941	1891	13 29
1889	859,412	1880	17.49	1901	1,554,157	1892	12.60
1890	903,833	1881	15.99	1902	1,639,606	1893	12 73
1891	936,846	1882	15.87	1903	1,713,684	1894	12 26
1892	993,307	1883	15.38	1904	1,825,634	1895	11.54
1893	1,032,492	1884	15.59	1905	1,946,496	1896	11.41
1894	1,062,876	1885	14.64	1906	2,064,566	1897	10 91
1895	1,090,922	1886	14.11	1907	2,210,147	1898	10 89
1896	1,127,468	1887	13 87	1908	2,348,493	1899	10 87

195. On donne le nom de « séries dynamiques irrégulières » à celles où se manifeste un mouvement alternatif d'avance et de recul, comme il arrive dans les séries relatives à des phénomènes qui traduisent l'état des relations

économiques. Ces séries peuvent être constituées de nombres exprimant les caractères de variabilité d'un phénomène unique, ou de nombres résultant de calculs ayant pour objet une pluralité de phénomènes ; on peut encore distinguer les séries dynamiques périodiques dans lesquelles se marque une tendance à la répétition des mêmes nombres aux mêmes époques.

EXEMPLE 9.

SÉRIES DYNAMIQUES IRRÉGULIÈRES				SÉRIES DYNAMIQUES PÉRIODIQUES				
BELGIQUE		BELGIQUE		BELGIQUE				
Exportations (Valeur du commerce spécial)		Nombre de grèves		Heures de soleil à Uccle 1886-1912				Jours sans soleil
Années	Milliers de frs.	Années	Nombres absolus	Mois	Etat possible	Total observé		Nombre moyen
						Moyen	P. c.	
1896	1,468	1896	139	Janvier. . .	263	60	23	13.1
1897	1,626	1897	130	Février. . .	280	82	29	8.6
1898	1,787	1898	91	Mars . . .	368	123	33	5.1
1899	1,949	1899	104	Avril . . .	415	176	42	2.3
1900	1,923	1900	146	Mai	481	222	46	1.6
1901	1,828	1901	117	Juin	493	207	42	1.7
1902	1,925	1902	73	Juillet . . .	496	226	46	1.1
1903	2,110	1903	70	Août. . . .	450	216	48	0.9
1904	2,183	1904	81	Septembre .	379	169	45	2.0
1905	2,334	1905	133	Octobre . .	332	123	37	4.9
1906	2,794	1906	207	Novembre. .	270	69	26	10.2
1907	2,848	1907	221	Décembre . .	247	53	21	12.8
1908	2,506	1908	101					
1909	2,810	1909	119	Année . . .	4,474	1,726	39	64.3
1910	3,407	1910	108					
1911	3,580	1911	156					
1912	3,951	1912	202					

II. — Sériation.

196. On appelle sériation l'opération par laquelle une masse de données résultant du calcul se trouve divisée en sections de manière à faire apparaître le nombre de cas rentrant dans chacune d'elles. La sériation a pour but principal d'aider à discerner les caractères essentiels des données statistiques en vue de tirer les conclusions scientifiques qui se dégagent des faits observés.

Chaque section opérée dans la série prend le nom de « classe »; la « grandeur » de la classe signifie l'étendue de la mesure dont on se sert à l'intérieur de la classe; si l'on observe la pression barométrique dans un endroit donné en la notant par mm., le millimètre est la grandeur de la classe. On donne le nom de « fréquence » à la quantité des cas observés appartenant à une classe donnée; l'ensemble des cas, répartis entre les diverses classes, et envisagés sous le rapport de leur répartition, est désigné sous le nom de « distribution des fréquences ». Le « module », selon la terminologie de certains auteurs, désigne la différence quantitative entre les termes extrêmes de la grandeur de deux classes successives (Bosco); dans une division de la population par âge, on commence parfois par énumérer cette population par mois pour la première année de vie, puis par année jusqu'à la cinquième année de vie, puis par cinq ans au delà, sauf à revenir à l'année pour les cas extrêmes de vieillesse; le module est successivement le mois, l'année, le lustre, et l'année. Toutefois, cette expression désigne, d'après la terminologie d'autres auteurs (Airy), la déviation-type (voir ch. III) multipliée par la racine carrée de 2 (1.4142136) (1). Pour éviter toute confusion, il est préférable de ne pas employer, dans la terminologie relative

(1) U. G. Yule : « An introduction to the theory of statistics », p. 144.

à la sériation, l'expression « module » dans le sens indiqué en premier lieu.

197. Le statisticien, en opérant la sériation des données fournies par le dépouillement, doit avoir en vue, d'une part, le but particulier qu'il se propose et, d'autre part, la nature intrinsèque du matériel qu'il met en œuvre.

La sériation étant un procédé d'analyse, sa formation doit toujours rester subordonnée au but que le statisticien se propose dans la recherche particulière qu'il entreprend; aussi plusieurs sériations différentes peuvent s'effectuer parmi les données statistiques en vue de faire apparaître différents caractères qu'elles possèdent. Quelques exemples sont ici nécessaires.

La population d'un pays, à chaque recensement, est présentée commune par commune; l'ensemble de ces données peut faire l'objet d'une sériation particulière en vue de reconnaître si une tendance se marque à la concentration dans les grandes villes. Pour trouver la solution d'un problème de ce genre, on peut avoir recours à une sériation d'après laquelle les communes du pays sont réparties en un certain nombre de catégories d'après leur population. Voici un tableau de ce genre relatif à la Belgique, d'après les données des neuf recensements généraux effectués dans ce pays; les données concernant l'année 1912 proviennent du relevé de la population d'après les registres de population; elles sont moins sûres que les suivantes, à raison des doubles emplois.

Exemple 10. — **BELGIQUE**. — **Répartition du nombre des communes en treize catégories, d'après la population.**

(*Annuaire statistique de la Belgique*, 1913, p. 49.)

Au 31 décembre des années	PROVINCES	100,000 habitants et au delà	50,000 à 100,000 habitants	25,000 à 50,000 habitants	20,000 à 25,000 habi ants	15,000 à 20,000 habitants	10,000 à 15,000 habitants	5,000 à 10.000 habitants	3,000 à 5,000 habitants	2,000 à 3,000 habitants	1,000 à 2,000 habitants	500 à 1,000 habitants	300 à 500 habitants	Moins de 300 habitants	Total des Communes
1912	Anvers	1	2	2	1	4	3	21	23	29	42	21	1	2	152
	Brabant	1	5	6	—	4	3	24	31	53	116	78	20	5	346
	Flandre Occidentale	—	1	3	1	2	5	38	40	33	71	36	11	9	250
	Flandre Orientale	1	—	2	2	2	10	32	52	51	80	51	13	2	298
	Hainaut	—	—	4	4	5	16	33	34	42	124	111	45	25	443
	Liége	1	—	2	1	1	7	19	28	27	87	93	50	26	342
	Limbourg	—	—	—	—	2	1	2	13	13	53	67	32	23	206
	Luxembourg	—	—	—	—	—	1	—	6	9	62	97	47	9	231
	Namur	—	—	1	—	—	—	7	11	10	58	141	77	59	364
	Le Royaume	4	8	20	9	20	46	176	238	267	693	695	296	160	2,632
1910	Le Royaume	4	7	20	10	18	45	169	243	265	697	701	300	150	2,629
1905		4	7	17	11	14	46	157	231	277	697	719	298	145	2,623
1900		4	6	15	7	14	42	139	230	270	705	734	306	145	2,617
1890		4	2	17	6	9	35	118	208	265	735	752	288	157	2,596
1880		4	—	14	5	9	30	104	209	254	737	760	296	161	2,583
1876		4	—	13	4	11	25	102	205	245	731	775	291	169	2,575
1866		3	1	5	7	9	19	88	181	260	719	772	306	181	2,551
1856		3	1	6	3	8	15	82	168	237	712	787	317	190	2,529
1846		2	2	4	5	3	15	81	163	249	694	756	339	208	2,521

Pour étudier la distribution des salaires dans une population ouvrière homogène, il faut opérer la sériation des salaires en adoptant une certaine grandeur, soit une division par franc si l'on se borne à des chiffres approximatifs, soit une division plus précise par cinquante ou vingt-cinq centimes.

Exemple 11. — **Salaires des hommes de plus de 16 ans (Belgique, industries des métaux, octobre 1903).**

TAUX DES SALAIRES	NOMBRE DES OUVRIERS		
	Sériation *a*	Sériation *b*	Sériation *c*
Total	84,136	84,136	84,136
Fr.			
1.50	978	(978)	(978)
1.50 à 1.74	1,495	2,937	(1,495)
1.75 à 1.99	1,442		4,141
2.00 à 2.24	2,699	4,962	
2.25 à 2.49	2,263		6,670
2.50 à 2.74	4,407	9,333	
2.75 à 2.99	4,926		13,974
3.00 à 3.24	9,048	16,127	
3.25 à 3.49	7,079		15,736
3.50 à 3.74	8,657	15,232	
3.75 à 3.99	6,575		14,646
4.00 à 4.24	8,071	12,616	
4.25 à 4.49	4,545		9,958
4.50 à 4.74	5,413	8,444	
4.75 à 4.99	3,031		7,184
5.00 à 5.24	4,153	5,859	
5.25 à 5.49	1,706		3,787
5.50 à 5.74	2,081	3,117	
5.75 à 5.99	1,036		2,403
6.00 à 6.24	1,367	1,938	
6.25 à 6.49	571		1,215
6.50 à 6.74	644	967	
6.75 à 6.99	323		(323)
7.00 et plus	1,626	(1,626)	(1,626)

La sériation *a* montre que la distribution des salaires à une échelle correspondant au quart du franc a une allure irrégulière : sauf dans un cas, le nombre des ouvriers gagnant un salaire compris entre le point du franc et le second échelon, ou entre le point du demi-franc et le quatrième échelon, est plus important que le nombre des ouvriers payés aux autres taux; l'équilibre se rétablit, et la série redevient régulière si l'on constitue des groupements de cinquante en cinquante centimes.

Lorsque la sériation est faite d'après des divisions étroites, correspondant aux limites entre lesquelles les faits sont observés, il arrive fréquemment que la série se présente sous une forme irrégulière, c'est-à-dire que les nombres se succèdent dans un ordre tantôt croissant, tantôt décroissant, au lieu d'augmenter d'une manière constante vers un maximum et de redescendre ensuite vers un minimum; on obvie à cet inconvénient en opérant un nouveau classement en partant du premier stade, puis du second et en comprenant dans une même division deux, et puis trois des divisions primitives. L'exemple précédent, emprunté à la statistique des salaires dans les industries des métaux, en Belgique (octobre 1903), montre d'une façon très claire le procédé à employer; nous le retrouverons plus loin en parlant de la façon empirique de rechercher la dominante.

198. La sériation par pour cent, au lieu d'utiliser les nombres absolus, peut aussi être employée et souvent elle donne des résultats plus instructifs que le simple groupement des données primaires. On peut calculer les pour cent en rapportant chaque partie à la masse, comme dans l'exemple suivant emprunté au tableau relatif à la répartition de la population belge dans les communes classées d'après leur grandeur :

Exemple 12.

	Communes de	Proportion du nombre de communes.
1910	100,000 habitants et plus	0.15
	50,000 à 100,000 habitants	0.30
	25,000 à 50,000 habitants	0.76
	20,000 à 25,000 habitants	0.34
	15,000 à 20,000 habitants	0.79
	10,000 à 15,000 habitants	1.78
	5,000 à 10,000 habitants	6.84
	3,000 à 5,000 habitants	9.00
	2,000 à 3,000 habitants	10.10
	1,000 à 2,000 habitants	26.14
	500 à 1,000 habitants	26.13
	300 à 500 habitants	11.07
	Moins de 300 habitants	6.00

On peut aussi rapporter les données primaires à une autre donnée plus générale avec laquelle elles accusent un rapport de dépendance, la population, par exemple, comme dans le tableau ci-après où la population active appartenant à l'industrie du vêtement est classée par province, en raison de la proportion par 1,000 habitants dans la province.

Exemple 13.

BELGIQUE : Recensement de l'industrie et du commerce, 1910

Brabant	45.2 p. m. de la population
Liége	28.9 p. m. de la population
Anvers	23.5 p. m. de la population
Flandre Orientale	23.3 p. m. de la population
Hainaut	22.4 p. m. de la population
Flandre Occidentale	21.7 p. m. de la population
Limbourg	20.4 p. m. de la population
Namur	20.2 p. m. de la population
Luxembourg	15.5 p. m. de la population
Le Royaume	27 5 p. m. de la population

199. Nous avons parlé, dans le chapitre consacré au dépouillement, de la grandeur des classes et du point de départ à adopter dans l'échelle du classement. (Cfr. n[os] 149 et 155.)

Ces observations s'appliquent également à la sériation. Celle-ci est d'ailleurs intimement liée au dépouillement en ce sens que les divisions de classes ne peuvent comporter une précision plus grande que celle admise pour le dépouillement lui-même. Si les salaires d'un groupe d'ouvriers ont été relevés et dépouillés en tenant compte seulement des divisions de cinquante en cinquante centimes, la sériation ne peut prétendre atteindre une plus grande précision, mais elle peut se faire en adoptant des bases plus larges, par exemple le franc au lieu de cinquante centimes. Cette façon de procéder est souvent utile, comme nous l'avons dit, en vue de faire disparaître des séries les inégalités accidentelles qui s'y remarquent et ainsi de mettre l'observateur sur la voie des tendances réelles qui se manifestent dans les groupes.

III. — Distribution des fréquences.

200. Après avoir disposé les séries, il est tout naturel d'examiner si la succession de leurs données obéit à une direction quelconque, si les « fréquences » augmentent ou diminuent progressivement ou si, après avoir suivi un mouvement ascensionnel, elles prennent ensuite une direction régressive. Cette étude paraît intimement liée à l'analyse des données statistiques. Supposons que l'on ait les chiffres de deux séries statistiques, concernant deux pays différents, mais relatives à la même période de temps; n'est-il pas nécessaire à l'étude comparative qu'on veut faire, de déterminer la direction de chacune de ces séries, le point où se manifeste un mouvement ascensionnel, la hauteur atteinte par ce mouvement, le degré où se marque l'arrêt et où commence la régression? Tous les lecteurs de statistiques se trouvent placés devant semblable problème; ils en cherchent la solution par des moyens plus ou moins simples, plus ou moins exacts. Or, il existe des procédés pour étudier et classer les courbes dessinées par les séries. C'est à l'exposé sommaire de cette matière, illustrée par quelques

données numériques et graphiques, qu'est consacré le présent paragraphe.

201. Pour analyser une série sous le rapport de la distribution de ses fréquences, il est souvent utile de déterminer graphiquement son développement. Dans ce but, on trace une première ligne horizontale, sur laquelle on porte, à intervalles égaux, les divisions des différentes classes dont se compose la série. Pour plus de facilité, on emploie du papier quadrillé ordinaire ou, ce qui est préférable mais plus coûteux, du papier millimétrique, avec les divisions de centimètre en centimètre plus marquées que les autres. On arrête, d'une façon arbitraire, la largeur à attribuer à chaque intervalle, par exemple 2 1/2 centimètres. Vers la gauche de cette ligne horizontale, on trace une seconde ligne, mais dans le sens vertical, recoupant la première à angle droit, de telle façon que le point de rencontre des deux droites soit l'origine ou point de départ du premier intervalle. Sur la ligne verticale on porte les divisions par fréquence en commençant par la plus basse et en finissant par la plus élevée; bien entendu, les intervalles sur l'échelle des fréquences sont proportionnels aux écarts des fréquences elles-mêmes. Pour chaque classe, on note alors, d'après l'échelle des fréquences, la hauteur correspondant à la fréquence de la classe.

Deux procédés s'offrent alors pour parfaire le graphique : relier par une suite de droites, formant une ligne brisée, les différents points déterminés comme ci-dessus. Si l'on adopte ce système, on doit indiquer le point de hauteur dans l'axe vertical du milieu de l'intervalle de la classe, commencer le tracé graphique à l'origine de l'intervalle, le faire passer par le point déterminé et de là le diriger sur le point indiquant la hauteur de la seconde classe, etc. On obtient alors ce qui s'appelle le *polygone de fréquence* de la série. Le second système consiste à élever, sur chaque intervalle des classes un rectangle ayant cet intervalle

comme base et dont la hauteur est proportionnelle à la fréquence de la classe : le professeur Pearson a donné à ces figures le nom de *histogramme.* Les deux figures sont représentatives, mais sous le rapport de l'exactitude le polygone de fréquence laisse quelque chose à désirer : veut-on mesurer l'aire de la figure, le polygone ne le permet qu'à la condition qu'il y ait compensation entre les angles formés par dessus et par dessous la ligne brisée, ce qui pratiquement ne se réalise que rarement. Aussi est-il préférable de recourir à l'emploi de l'histogramme; cette figure permet de mesurer exactement l'aire du phénomène exprimé au moyen de la série. Cependant, lorsqu'il s'agit de figures comparatives, le choix doit être donné au polygone de fréquence; les lignes superposées dans l'histogramme double servant à une comparaison sont embrouillées et bien moins parlantes que les courbes plus ou moins élancées ou plus ou moins aplaties des polygones comparatifs (1).

202. Une première hypothèse semble surgir naturellement quand on en vient à rechercher la forme de la courbe décrite par les données relatives à certains phénomènes statistiques : pourquoi cette courbe ne ressemblerait-elle pas à la courbe normale des erreurs, à la courbe de Gauss, ou à la courbe que Quetelet a popularisée, au moins parmi les statisticiens, sous la dénomination de courbe binomiale? Nous n'avons pas dit que la loi de distribution serait *la même,* identique à la loi de Gauss, mais seulement qu'elle pourrait y ressembler, car, en général, il est avéré que la loi de la distribution normale des erreurs n'est fondée que pour les mesures provenant d'observations répétées d'un même objet (2).

---

(1) Voir l'exposé de la statistique graphique, livre II, chap. V.

(2) Il n'est pas inutile d'insister sur ce point de vue, car de vives controverses se sont élevées à ce sujet. Il y a quelque vingt ans, le professeur Pearson a critiqué avec vivacité ceux de ses prédécesseurs qui avaient établi, selon lui, une analogie trop étroite entre la loi des erreurs accidentelles et la distribution des faits statistiques. La formule de la loi des erreurs donne

Lorsque la loi définie par Gauss, et qui originairement avait été trouvée par Laplace, fut reconnue et appliquée, on la trouva exacte pour les mesures différentes prises sur un même objet : ces mesures différentes s'appliquaient à des matières qui sont le sujet ordinaire des observations dans les sciences astronomiques, physiques, anthropologiques, etc. S'agit-il, comme dans les statistiques, non plus d'un objet, mais d'une série d'objets différents quoique homogènes, la forme de la distribution semble être, en général, asymétrique autour de la moyenne. Ceci n'empêche pas qu'une ressemblance pourrait exister et qu'il y a un très grand intérêt à la rechercher et à la déterminer.

Il n'entre pas dans notre plan de discuter à cette place la loi des erreurs accidentelles dont on trouvera un exposé au livre III; mais il est cependant impossible de résumer d'une manière intelligible ce qui a trait à la distribution des fréquences sans faire précéder ces indications de quelques notions préliminaires sur la loi des erreurs.

D'après Gauss, la loi des erreurs, qui est traduite graphiquement par la courbe normale des erreurs ou courbe normale des fréquences, exprime la probabilité comparative des erreurs de différente grandeur. Le raisonnement, autant que l'expérience, nous avertissent suffisamment de ce que, lorsqu'un observateur n'est guidé par aucun parti pris et qu'il désire réellement observer ce qui est placé sous ses yeux (*animus observandi*), les erreurs les plus considérables sont aussi les moins nombreuses, tandis qu'il y a des chances pour que les petites erreurs soient beaucoup plus fréquentes. Un second principe démontré par Gauss est que les erreurs positives, c'est-à-dire en trop, et les erreurs

---

naissance à une courbe qui, en fait, est éloignée de la distribution des phénomènes considérés en statistique; les travaux du professeur Pearson ont jeté sur la question une vive lumière. On en trouvera ci-après un aperçu, nécessairement très simplifié et très réduit, car il ne pouvait être question d'aborder dans cet ouvrage, qui s'efforce d'être utile à tous, l'appareil mathématique compliqué dont les démonstrations de l'espèce sont nécessairement entourées.

négatives, autrement dites erreurs en moins, sont également probables.

La démonstration de Gauss, sur laquelle nous n'avons pas à nous arrêter ici, est purement mathématique. Son principe est évidemment juste ; cependant, un logicien-économiste de grande valeur, Stanley Jevons, a fait observer que le fondement du raisonnement de Gauss repose, somme toute, plutôt sur une hypothèse plausible que sur la preuve d'une véritable nécessité, d'une loi au sens propre (1).

Aussi le savant anglais fait-il remarquer qu'on peut arriver à la même démonstration par une tout autre voie, qui est celle à laquelle Laplace et Quetelet ont eu recours. Un statisticien autrichien, M. le Dr Zizek, caractérise ainsi la méthode de Quetelet : « Quetelet a trouvé qu'il existe une distribution symétrique autour de la moyenne, correspondant à la loi des erreurs, spécialement dans les séries relatives à la taille et à l'ampleur de la poitrine. Dans ses recherches, Quetelet ne fait pas usage de la loi de Gauss, mais de la formule binomiale, dans laquelle les probabilités d'écart de la moyenne correspondent aux coefficients d'expansion binomiale (2). » On peut donc faire emploi du triangle arithmétique pour déterminer la distribution de la courbe symétrique, *quand on connaît le nombre des erreurs*. Ainsi si nous admettons que dans une recherche il y ait six chances d'erreur positive et autant d'erreur négative, nous consultons la treizième ligne du triangle arithmétique et nous écrivons, en suivant le modèle donné par Stanley Jevons (3), les résultats comme suit :

Direction des erreurs	Erreurs positives		Erreurs négatives
Importance des erreurs. .	6, 5, 4, 3, 2, 1	0	1, 2, 3, 4, 5, 6
Nombre de chaque erreur.	1, 12, 66, 220, 495, 792	924	792, 495, 220, 66, 12, 1

(1) Stanley Jevons, *The principles of Science*, p. 378.
(2) Zizek, *Statistical averages*. New-York, 1913, p. 276.
(3) Stanley Jevons, *loc. cit.*, p. 379.

De quoi il résulte, comme le montre encore le logicien anglais, que la probabilité d'une erreur positive ou négative est de 792/4096, fraction dans laquelle le numérateur est le nombre de combinaisons donnant une erreur du degré le plus faible (on suppose connue la grandeur des erreurs) et le dénominateur représente le nombre de toutes les erreurs de toute grandeur. De même, nous pouvons aisément déterminer la portion de la courbe affectée d'une erreur d'une certaine importance : par exemple, le troisième degré. Il suffit de noter les chiffres placés sous les numéros inscrits à la seconde ligne : 3, 2, 1, 0, 1, 2, 3 et de les additionner ; on a ainsi 3938/4096, qui exprime la probabilité de réunir dans un total toutes les erreurs positives et négatives dont l'importance ne dépasse pas le troisième degré.

Mais d'habitude on ne connaît ni le nombre des erreurs possibles, ni leur grandeur, de telle sorte que les mathématiciens raisonnent comme si l'on avait un nombre infini d'erreurs, hypothèse fort plausible qui a l'avantage de fournir une loi générale applicable, par la voie du calcul, à tous les problèmes. La courbe formée d'après la formule générale (1) basée sur un nombre infini d'erreurs et celle tracée en supposant qu'il n'y ait qu'un petit nombre de causes d'erreurs, se ressemblant au point qu'on les distinguerait difficilement l'une de l'autre.

Quetelet ayant adopté cette démonstration par les coefficients d'expansion binomiale, qui avait été utilisée avant lui, ainsi du reste que les échelles de possibilité et de précision qui en dérivent, eut l'idée de l'appliquer à certaines mesures anthropométriques. Il ne semble pas que cette idée lui vint avant 1843 (2), mais dès qu'elle fut née il la déve-

---

(1) Cette formule générale est (d'après Stanley Jevons) : $y = Ye^{-cx^2}$, dans laquelle x représente l'importance de l'erreur, Y est l'ordonnée maximum de la courbe, C est un nombre constant exprimant l'importance de la tendance à l'erreur entre chaque série d'observations. e est une constante mathématique exprimant la base des logarithmes hyperboliques (2 — 7182818). (Stanley Jevons, *loc. cit.*, p. 381.)

(2) Cfr. Lottin, *Quetelet*, Louvain, 1912, pp. 151 et suiv.

loppa avec une complaisance singulière, notamment dans ses fameuses « Lettres sur la théorie des probabilités » publiées en 1846, et il fut même tenté de lui attribuer une portée que des travaux ultérieurs, dus à d'autres statisticiens, ne permettent décidément pas de lui accorder.

203. Après cette esquisse rapide des préliminaires de la question, revenons à notre point de départ et demandons-nous quel est le rapport qui existe entre la courbe normale des erreurs, la courbe binomiale et la répartition des fréquences entre les classes d'un phénomène relevé par la statistique. Ce rapport, qui avait d'abord paru très fréquent, est, en réalité, assez rare, si l'on veut parler de courbes ayant une forme très rapprochée de la courbe normale. Il n'est pas douteux qu'elle n'est pas une loi générale de distribution des phénomènes statistiques. En dehors de l'anthropométrie, on n'en trouve d'application que dans des cas exceptionnels; même dans le domaine de l'anthropométrie, il se présente bien des phénomènes dont la loi de distribution est asymétrique (1). Dans quelques cas, la régularité de la courbe des erreurs est reproduite par la courbe de fréquence. « Entre le polygone binomial et la courbe normale de fréquence, dit M. Pearson, il existe une relation très étroite qui est de nature géométrique et qui est presque indépendante de la grandeur des *n* (2). » Pratiquement, la réalisation parfaite de cette courbe est un cas qui ne se présente que fort rarement; on en obtient le plus souvent seulement des images plus ou moins fidèles qui reproduisent l'allure générale de la courbe sans en présenter toute la rigueur mathématique.

(1) La distribution des fréquences en ce qui concerne le *poids* des individus est du nombre : While the distribution of stature is in general symmetrical, that of weight is asymmetrical or *skew*, the greater frequencies lying towards the lower end of the range. (YULE, *An Introduction to the theory of statistics*, p. 93, et ZIZEK, *loc. cit.*, p. 277.)

(2) PEARSON, « Contributions to the mathematical theory of evolution. II Skew variation in homogeneous material ». *Philosophical transactions of the Royal Society of London*, A, vol. 186 (1895).

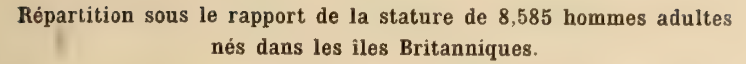
Le lecteur trouvera, ci-après, suivi du diagramme correspondant, un tableau indiquant la taille de 8,585 personnes nées dans les îles Britanniques. Ce tableau, reproduit d'après M. Yule, qui l'a emprunté au « Final report » de la commission anthropométrique de la « British Association » (1883, p. 256), présente une répartition des fréquences sensiblement égale à la loi normale des erreurs.

Exemple 14.

Stature d'hommes adultes nés en Angleterre, en Ecosse, dans le pays de Galles et en Irlande.

Taille sans souliers en pouces.	Nombre d'hommes compris dans les dites limites de grandeur. Endroit de naissance.				TOTAL.
	Angleterre	Ecosse.	Galles.	Irlande.	
57	1	—	1	—	2
58	3	1	—	—	4
59	12	—	1	1	14
60	39	2	—	—	41
61	70	2	9	2	83
62	128	9	30	2	169
63	320	19	48	7	394
64	524	47	83	15	669
65	740	109	108	33	990
66	881	139	145	58	1223
67	918	210	128	73	1329
68	886	210	72	62	1230
69	753	218	52	40	1063
70	473	115	33	25	646
71	254	102	21	15	392
72	117	69	6	10	202
73	48	26	2	3	79
74	16	15	1	—	32
75	9	6	1	—	16
76	1	4	—	—	5
77	1	1	—	—	2
Totaux	6194	1304	741	346	8585

(*Voir diagramme, page suivante, fig.* 15.)



Répartition sous le rapport de la stature de 8,585 hommes adultes nés dans les îles Britanniques.

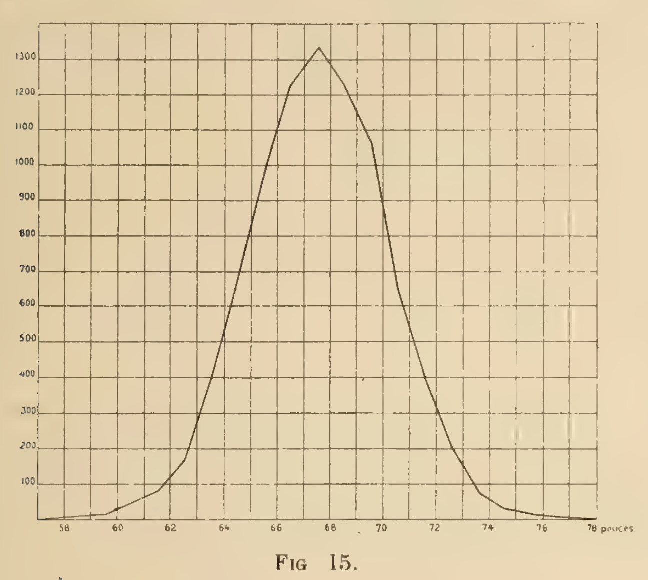


Fig 15.

204. Si les séries présentant un développement absolument régulier sont rares, elles sont par contre fort nombreuses, celles dont la courbe s'écarte légèrement de la normale, mais dont la disposition générale rappelle la forme de la loi des erreurs. Quetelet ne s'était pas préoccupé de différencier les diverses espèces de courbes, il s'en était tenu à la démonstration générale de la loi binomiale. C'est à l'école contemporaine anglaise des statisticiens-mathématiciens que revient le mérite d'avoir dégagé les formes usuelles des courbes statistiques et d'en avoir indiqué les formules.

Plusieurs méthodes se partagent les mathématiciens de la statistique en ce qui concerne la manière d'identifier et de calculer les courbes dont il s'agit. Une première méthode part de ce point de vue général que toutes les courbes de distribution par fréquence sont originairement dérivées de la courbe normale des erreurs; qu'il importe donc que les

formules employées pour le calcul des courbes se rapprochent le plus possible de la formule de la courbe normale des erreurs, et que si même la dite formule ne s'applique pas entièrement au matériel statistique, elle est encore préférable à une formule empirique plus exacte, parce que celle-ci ne pourrait fournir aucune explication de la forme de distribution envisagée (1).

Le professeur Edgeworth (2) a exprimé l'opinion que les courbes asymétriques sont probablement très fréquentes dans le domaine des sciences naturelles et anthropologiques; c'est ainsi qu'au lieu d'une courbe normale de la stature en Italie, on peut tracer un certain nombre de courbes qui correspondent à des types réellement distincts d'après les différentes provinces. Dans un autre domaine, M. Pearson s'est livré à une décomposition fort curieuse de la courbe de mortalité qu'il a divisée en ses principaux éléments et il a fait porter ses calculs sur les Anglais du sexe masculin et sur la population française des deux sexes. La concordance presque complète des résultats dans les deux cas, dit M. Pearson, lui a inspiré confiance dans l'exactitude des résultats. Les cinq éléments fondamentaux

---

(1) Se basant sur ces principes, le professeur Pearson a recherché, dans une suite de travaux remarquables, les formules les plus adéquates pour diviser en deux ou trois courbes normales la représentation des séries qui ne paraissent pas obéir à la courbe normale des erreurs. Antérieurement au savant anglais, M. Bertillon père avait déjà montré que si deux types d'hommes se trouvaient juxtaposés dans un pays, la présence des deux races se trahirait dans le diagramme des tailles; s'inspirant des considérations développées par Quetelet, le docteur Bertillon père a en effet montré un exemple fort remarquable d'une courbe à deux sommets relative à la taille des conscrits dans le département du Doubs (1858-1867). La méthode de M. Pearson pour diviser en plusieurs courbes régulières une courbe asymétrique a été, au point de vue mathématique, l'objet de grands éloges; on lui a cependant reproché d'être très laborieuse et l'on a suggéré quelques simplifications.

(2) « On the representation of statistics by mathematical Formulæ » (part. II), par le professeur Edgeworth. (*Journal of the Royal Statistical Society*, 1899, p. 126.)

de la courbe de mortalité pour les Anglais du sexe masculin sont les suivants, les nombres se rapportant à 1,000 personnes nées dans la même année (1) :

Mortalité de la vieillesse, verticale-centrale ou dominante (mode) (2), 67 ans;

Mortalité de l'âge mûr, verticale-centrale ou dominante, 41.5 ans;

Mortalité de la jeunesse, verticale-centrale ou dominante, 22.5 ans;

Mortalité de l'enfance, verticale-centrale ou dominante, 6.06 ans;

Mortalité du premier âge, verticale-centrale ou dominante, 156 jours.

205. Une seconde méthode est celle préconisée par Fechner dès 1878 et développée par lui en 1897, on pourrait l'appeler la « méthode de reconstitution »; elle est basée sur l'hypothèse que la courbe de Gauss pourrait, dans certains cas, être asymétrique.

L'auteur de cette méthode a été frappé d'une part de la ressemblance qu'il y a entre la courbe normale des erreurs et la distribution des fréquences dans de nombreuses séries statistiques; d'autre part, il n'a pas manqué de remarquer que les cas de similitude complète étaient fort rares. De là une hypothèse, à première vue assez plausible, a surgi : les courbes irrégulières ne seraient-elles pas deux parties de la courbe normale mais dont le degré de dispersion serait différent ? Quetelet, sans procéder à une analyse mathématique, avait, somme toute, indiqué les données de cette hypothèse en développant les conséquences de la loi binomiale;

---

(1) PEARSON, *Contributions to the mathematical theory of evolution*, p. 407.

(2) La dominante ou *mode* est la mesure prise sur l'axe horizontal (axe des abscisses) qui correspond à la hauteur la plus grande prise sur l'axe vertical (axe des ordonnées). Pour les données élémentaires de la construction graphique, cfr. *supra* n° 201.

il commence par tracer la courbe de possibilité pour un tirage de boules de deux couleurs de 999 boules ; cette courbe a la forme d'une cloche aux parois très évasées ; s'il s'agit de calculer et de construire la courbe de possibilité pour un nombre de boules quadruple (3996) et la perpendiculaire (le mode) représentant la probabilité de l'événement le plus probable, on obtient une courbe en forme de cloche beaucoup plus resserrée et qui ressemble à la pointe d'une toupie ; finalement, « plus le nombre de boules que l'on prend à chaque tirage est considérable, plus il tend à s'établir une égalité entre le nombre de boules blanches et le nombre de boules noires. Si les choses étaient poussées à l'extrême, et si chaque tirage amenait un nombre infini de boules, la ligne de possibilité se réduirait insensiblement à la perpendiculaire... (1) ». Or, si l'on réunissait deux moitiés de courbes, l'une se rapportant à un nombre x de boules et l'autre à un nombre différent, on obtiendrait une courbe irrégulière se rapprochant de la normale et dont l'irrégularité serait due au nombre différent d'épreuves exprimé par chaque partie.

La méthode de Fechner consiste donc à considérer séparément chaque moitié d'une courbe asymétrique et à la compléter par une autre moitié identique à la première, de façon à construire une courbe plus ou moins évasée, mais régulière. Lorsque l'écart de l'extrémité de la courbe au centre n'est pas trop grand, on peut, d'après Fechner, procéder en opérant sur les données arithmétiques, mais si l'écart est très considérable, il est préférable, selon cet auteur, de substituer aux nombres eux-mêmes leurs logarithmes, de placer ces logarithmes en série, de chercher la dominante (mode) et d'étudier les écarts des logarithmes autour de leur mode. La raison en est que la déviation purement arithmétique indique seulement la différence entre la

---

(1) QUETELET, *Lettres sur la théorie des probabilités*. Bruxelles, 1846, lettre XVI, pp. 106 (fig.) et 107.

moyenne et les nombres naturels, tandis que les mesures prises sur les logarithmes indiquent les variations relatives, d'après lesquelles on peut mieux se rendre compte des changements survenus dans les phénomènes collectifs.

206. D'autres méthodes d'analyse des courbes ont encore été essayées par les statisticiens-mathématiciens, notamment par Edgeworth. La classification générale des types de courbes asymétriques a été proposée par le professeur Pearson sur la base de cinq types (1), auxquels il a plus tard ajouté un sixième type en U (2). Les six types reconnus et calculés par M. Pearson sont les suivants :

Type I. Axe des abscisses limité de part et d'autre, courbe asymétrique.

Type II. Axe des abscisses limité de part et d'autre, symétrique.

Type III. Axe des abscisses limité d'un seul côté, courbe asymétrique.

Type IV. Axe des abscisses illimité des deux côtés, courbe asymétrique.

Type V. Axe des abscisses illimité des deux côtés, courbe symétrique ou courbe normale des erreurs, de Gauss, dont nous avons parlé plus haut (voir n° 202).

Type VI. Forme de la courbe en U.

Pour déterminer le type de courbe auquel se rapporte une série, il faut procéder à des calculs compliqués et laborieux. On doit se demander en premier lieu si l'on se trouve devant une courbe comprenant un seul « mode » ou en comprenant plusieurs ; ce point déterminé, on a à rechercher si

---

(1) PEARSON, « Contribution to the theory of mathematical evolution. Skew variations in homogoneous material ». *Philosophical Transactions of the Royal Society of London*, 1895, I, p. 360.

(2) « Cloudiness » : Note on a novel case of Frequency, by Karl PEARSON. (*Proceedings of the Royal Society*, LXII, 1897-1898, p. 287 à 290.)

la courbe est simple ou complexe; les courbes étudiées par le professeur Pearson, et que nous venons d'énumérer, sont des courbes simples. En général, les courbes qui ne comprennent qu'un seul « mode » et sont basées sur un matériel homogène, appartiennent à la classe des courbes simples. On peut donc, dans un cas semblable, confronter les graphiques à ceux représentant les courbes typiques de Pearson et les classer dans l'une ou l'autre catégorie sans devoir se livrer à des calculs laborieux. L'exposé des méthodes mathématiques du professeur Pearson ne peut se faire à cet endroit, parce qu'il suppose connus des points dont l'examen n'est abordé que plus loin (détermination de la moyenne, de la dominante et de la médiane, de la standard déviation, etc.). On en trouvera un exposé au chapitre II de notre livre III. Nous nous bornerons ici à reproduire des exemples des différents types de courbes et à y joindre le graphique correspondant.

207. Pour plus de commodité, nous réunirons dans cet exposé plusieurs types présentant des caractères à peu près semblables et, en premier lieu, nous en examinerons les courbes de distribution légèrement asymétriques dans lesquelles la distribution par fréquence croît et décroît plus rapidement d'un côté que de l'autre. Ces courbes sont les plus fréquentes; on en trouve des exemples dans tous les domaines et elles abondent dans les questions économiques. C'est ainsi que M. Yule a trouvé un excellent exemple de distribution modérément symétrique dans un tableau donnant pour l'Angleterre et le pays de Galles, en 1891, le nombre de districts d'enregistrement comptant un certain pourcentage de la population secourue (1).

---

(1) U. G. YULE, *An introduction to the Theory of Statistics*, London, 1911, pp. 92-93. Cfr. *Analyse mathématique de la courbe*, par le prof. PEARSON, *loc. cit.*, p. 405.

Nous reproduisons ci-après le tableau dressé par M. Yule et le polygone de fréquence qui en traduit les données (*voir diagramme page suivante, fig. 16*) :

EXEMPLE 15.

Proportion de la population recevant des secours.	Nombre d'unions comptant la proportion ci-contre.
0.75 — 1.25	18
1.25 — 1.75	48
1.75 — 2.25	72
2.25 — 2.75	89
2.75 — 3.25	100
3.25 — 3.75	90
3.75 — 4.25	75
4.25 — 4.75	60
4.75 — 5.25	40
5.25 — 5.75	21
5.75 — 6.25	11
6.25 — 6.75	5
6.75 — 7.25	1
7.25 — 7.75	1
7.75 — 8.25	0
8.25 — 8.75	1
TOTAL.	632

La distribution des salaires parmi une population ouvrière homogène fournit aussi un bon exemple de courbe légèrement asymétrique. On y voit nettement qu'un assez grand nombre d'hommes atteignent rapidement le taux auquel correspond le « mode »; puis, vers les taux supérieurs, leur nombre diminue sensiblement; vers la fin de la série, les proportions d'ouvriers gagnant les plus hauts salaires deviennent très minimes (*fig. 17*).

Proportion de la population secourue. — Angleterre, 1891.

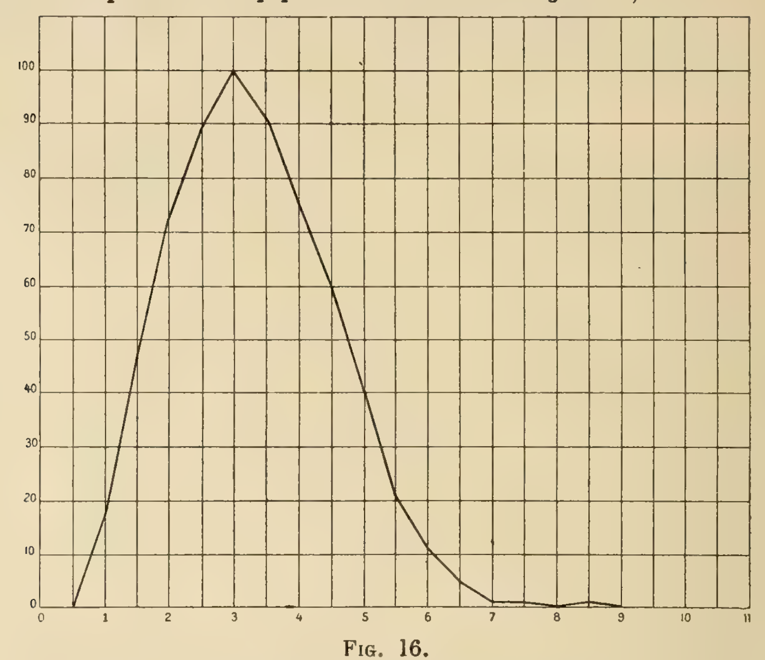


Fig. 16.

Distribution des salaires parmi les ouvriers des mines de houille travaillant au fond. — Belgique 1896.

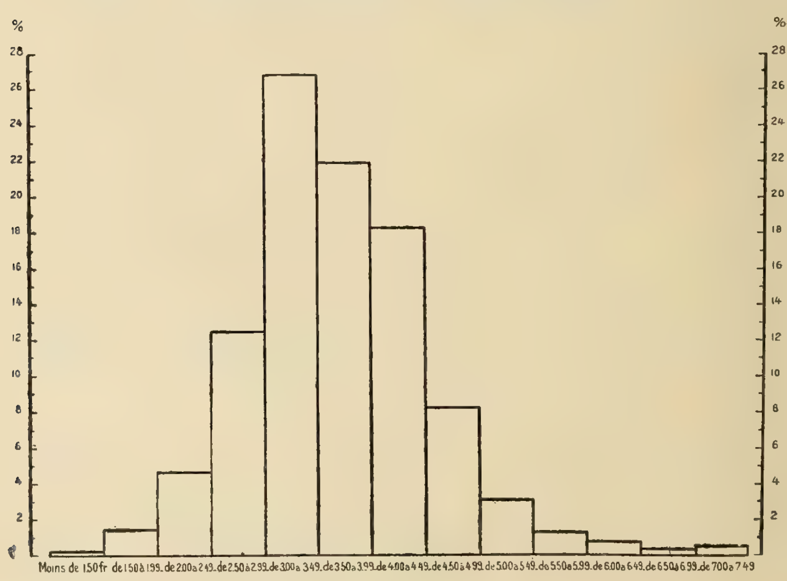


Fig. 17.

BELGIQUE. — Taux des salaires dans les mines de houille, octobre 1896 (1).

EXEMPLE 16.

Taux des salaires.	Nombre d'hommes adultes du fond ayant touché les salaires ci-contre.
Moins de fr. 1.50	155
de 1.50 à 1.99	861
2.00 à 2.49	2,860
2.50 à 2.99	7,660
3.00 à 3.49	16,456
3.50 à 3.99	13,444
4.00 à 4.49	11,235
4.50 à 4.99	5,057
5.00 à 5 49	1,888
5.50 à 5.99	785
6.00 à 6.49	439
6.50 à 6.99	190
7.00 à 7.49	263
TOTAL.	61,293 (2)

Lorsqu'il y a une augmentation générale des salaires, la courbe se déforme : elle s'aplatit, le nombre de classes augmente vers la droite et les fréquences y sont plus élevées que précédemment.

Pour emprunter un exemple à la physiologie, nous reproduisons ci-après, d'après Quetelet, qui en a fait usage dans ses « Lettres sur la théorie des probabilités », un tableau donnant les mesures prises sur les poitrines de 5,738 soldats des régiments écossais. Quetelet avait utilisé les chiffres publiés dans le treizième volume du *Journal médical d'Edimbourg* (3).

(1) Chiffres extraits de la *Statistique des salaires dans les mines de houille,* octobre 1896-mai 1900, Office du Travail, Bruxelles, 1901, p. 34.

(2) Il n'a pas été tenu compte de six ouvriers occupés dans un charbonnage à un travail spécial et ayant gagné des salaires supérieurs à fr. 7.50.

(3) QUETELET, *Lettres sur la théorie des probabilités.* Bruxelles, 1846, p. 136.

Exemple 17.

Mesures de la circonférence des poitrines des soldats écossais.

Mesures de la poitrine (pouces).	Nombre d'hommes.
33	3
34	18
35	81
36	185
37	420
38	749
39	1,073
40	1,079
41	934
42	658
43	370
44	92
45	50
46	21
47	4
48	1
Total	5,738

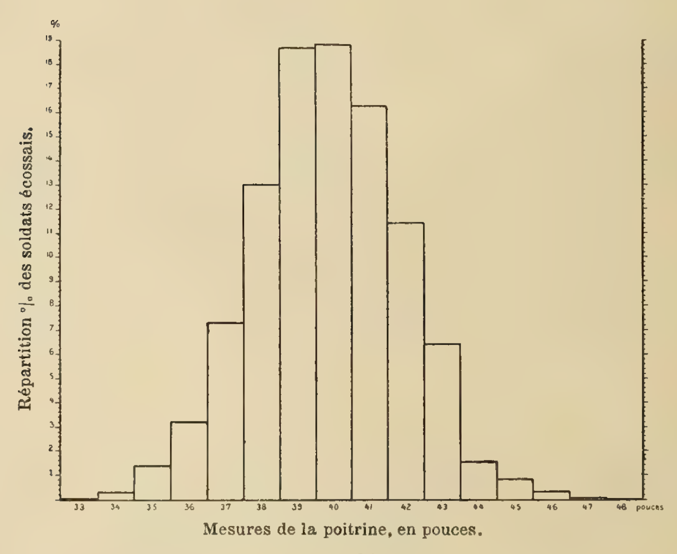


Fig. 18.

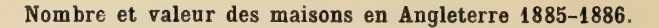
208. *Formes de distribution entièrement asymétriques.* — Lorsque la courbe présente un maximum dès le début, suivi d'une chute rapide, la courbe est dite asymétrique. C'est la forme qu'affecte la courbe des revenus calculée par Vilfredo Pareto, forme bien connue et qui ressemble à la pointe d'une toupie qu'on aurait coupée en deux par le milieu. La même forme se retrouve dans un grand nombre de phénomènes économiques et dans la répartition de certains faits naturels, principalement en botanique. Il importe cependant de faire une réserve : peut-être la forme asymétrique de la courbe n'est-elle due qu'à l'insuffisance numérique des observations ou au fait qu'elles n'ont pas été faites d'une manière assez précise. Parmi les faits économiques, l'un de ceux qui fournissent une bonne illustration est la répartition de la valeur fiscale des maisons, avec le nombre de celles-ci, en Angleterre et dans le Pays de Galles. M. Pearson, dans son mémoire déjà cité, a analysé les chiffres produits par M. Goschen, dans un discours présidentiel adressé à la Société de Statistique de Londres en 1887 et a montré que les résultats du calcul et de la statistique ne s'écartent l'un de l'autre que dans une proportion très faible (1).

Voici les chiffres de Goschen; nous les faisons suivre de l'histogramme calculé par M. Pearson (*voir diagramme page suivante, fig. 19.*)

Exemple 18. — **Maisons; Angleterre et Pays de Galles, 1885-1886.**

Taux d'évaluation.	Nombre de maisons.	Taux d'évaluation.	Nombre de maisons.
Moins de 10 £.	3,174,806	£ 80 à £ 100	47,326
£ 10 à £ 20 .	1,450,781	£ 100 à £ 150	58,871
£ 20 à £ 30 .	441,595	£ 150 à £ 300	37,988
£ 30 à £ 40 .	259,756	£ 300 à £ 500	8,781
£ 40 à £ 50 .	150,968	£ 500 à £ 1000	3,002
£ 50 à £ 60 .	90,432	£ 1000 à £ 1500	1,036
£ 60 à £ 80 .	104,128		

(1) Pearson, *loc. cit.*, pp. 396 et suivantes.


Nombre et valeur des maisons en Angleterre 1885-1886.

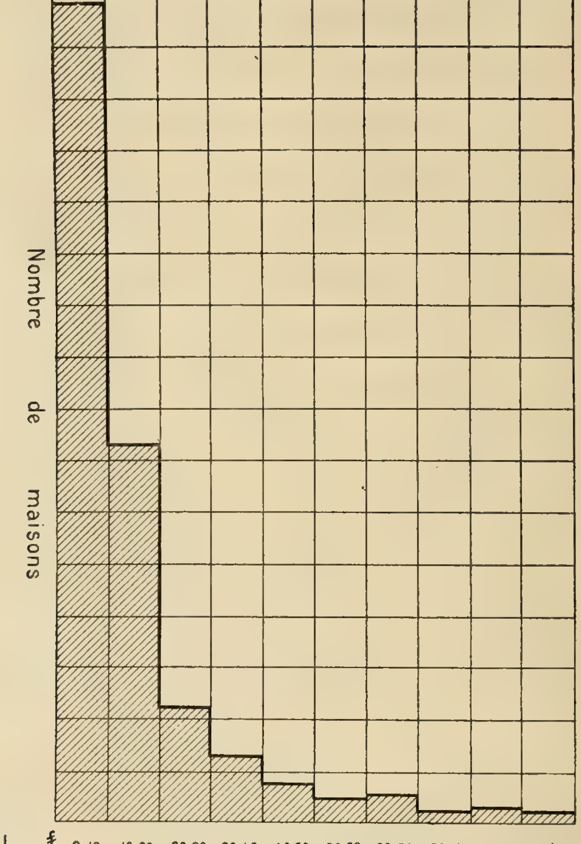


Fig. 19

Des courbes analogues se rencontrent, avons-nous déjà dit, à propos de phénomènes d'ordre botanique. La variation du nombre de pétales des fleurs d'une espèce donnée a souvent donné lieu à de pareilles constatations. En 1887, M. H. De Vries ayant repiqué dans son jardin un certain nombre d'exemplaires de *Ranunculus bulbosus,* compta l'année suivante le nombre de pétales que présentaient 222 fleurs de ces plantes. Les résultats sont consignés ci-après :

Nombre de pétales	5	6	7	8	9	10
Fréquence	133	55	23	7	2	2

Ici encore, selon M. Pearson, les résultats du calcul et de l'expérience s'accordent d'une manière remarquable (1).

209. *Formes de distribution asymétriques en U.* — Après avoir exposé les formes de distribution indiquées ci-dessus, M. Pearson, dans son mémoire sur les *Skew variations,* disait que, théoriquement, on pouvait s'attendre à rencontrer des cas de distribution des fréquences se trouvant à l'opposite de la distribution normale, c'est-à-dire que les extrémités de la courbe correspondraient à un maximum, tandis que le minimum de fréquence se trouverait au centre. En 1897, M. Pearson communiqua à la « Royal Society » de Londres un exemple de cette forme spéciale de distribution qu'il appela forme en U (2). Cet exemple est emprunté à un relevé, divisé en 10 degrés, du point de nébulosité (cloudiness) à Breslau pendant les années 1876-1885, comprenant 3,653 jours d'observation.

Les données numériques de la courbe sont les suivantes :

Degrés	0	1	2	3	4	5	6	7	8	9	10
Fréquence	751	179	107	69	46	9	21	71	194	117	2089

Bien que la distinction par degrés de nébulosité durant une journée entière renferme nécessairement une part d'approximation, M. Pearson note cependant que le rapport entre le calcul et l'observation est, dans ce cas encore, fort satisfaisant.

On peut rapprocher de cet exemple les données suivantes qui concernent le nombre moyen de journées sans soleil, à l'observatoire royal d'Uccle, pendant les années 1886 à 1912. Voici les données publiées à ce sujet par l'Institut

(1) PEARSON, *loc. cit.*, p. 401.

(2) PEARSON, « Cloudiness » : Note on a novel case of Frequency. (*Proceedings of the Royal Society*, vol. LXII (1897), p. 287.)

royal météorologique (1), suivies de l'histogramme qui s'y rapporte (*voir diagramme ci-dessous, fig. 20*).

EXEMPLE 19.

Mois.	Nombre moyen de jours.	Mois.	Nombre moyen de jours.
Janvier	13,1	Juillet	1,1
Février	8,6	Août	0,9
Mars	5,1	Septembre	2,0
Avril	2,3	Octobre	4,9
Mai	1,6	Novembre	10,2
Juin	1,7	Décembre	12,8

Nombre moyen de jours sans soleil. — Uccle, 1886-1912.

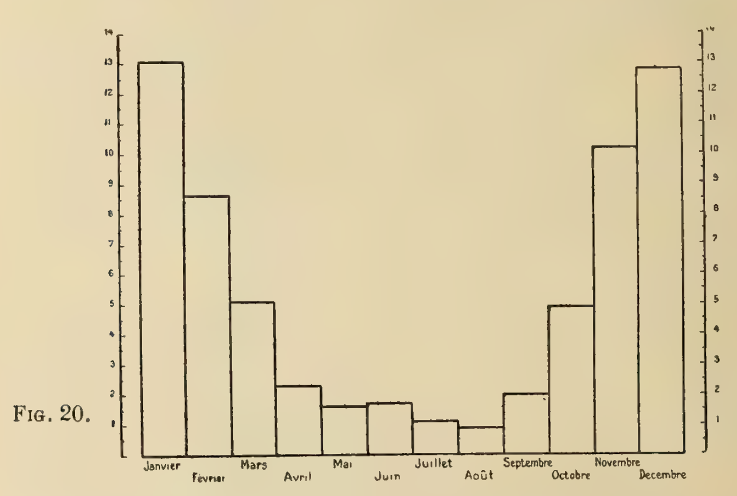


FIG. 20.

210. Pour terminer cet exposé sommaire, il reste à se demander quel est le résultat scientifique obtenu par les calculs auxquels nous avons fait allusion plus d'une fois. Il faut distinguer, semble-t-il, entre les formules purement empiriques et celles qui sont basées sur la loi de distribution des erreurs. Les formules purement empiriques peuvent être considérées comme une représentation mathématique d'un phénomène, mais elles n'expriment pas la loi du phénomène.

Il en va autrement pour les formules basées sur la loi

(1) Belgique. *Annuaire météorologique*, 1914, p. 25.

de Gauss. Edgeworth a fait observer qu'elles fournissent une explication plausible de la construction des séries auxquelles elles s'appliquent. Lorsque la distribution normale des erreurs concerne des mesures prises un grand nombre de fois sur un même objet, on doit supposer que l'on se trouve en présence d'un grand nombre d'erreurs accidentelles, positives et négatives, qui se contrebalancent et sont d'autant moins nombreuses qu'elles sont plus importantes; c'est l'hypothèse fondamentale de Gauss. Par assimilation, on a étendu la loi de Gauss à des observations nombreuses portant sur des objets différents, mais de même nature, et l'on a trouvé que, dans certains cas, la loi normale des erreurs pouvait aussi trouver son application. Mais ces cas, — on l'a reconnu après s'être laissé trop facilement suggestionner par cette hypothèse séduisante — sont, somme toute, fort rares; on ne les rencontre guère que dans le domaine des recherches anthropologiques et pour un nombre de faits limités seulement. Ce qui est le cas le plus fréquent, ce sont les courbes irrégulières qui rappellent la courbe normale de Gauss, mais s'en éloignent par quelque point. Mais lorsqu'on observe une courbe irrégulière, on ne peut plus dire qu'on se trouve en présence d'erreurs accidentelles commises en nombre égal dans le sens positif comme dans le sens négatif et dont l'importance diminue à mesure qu'elles s'écartent de la moyenne. La seule explication plausible reste alors que des forces opposées et d'importance inégale luttent en sens contraire. L'infinie variété de ces courbes nous écarte alors de la notion simple de loi. Si la distribution des salaires, à une époque donnée, s'exprime par une certaine courbe, et quelques années plus tard par une courbe toute différente (1), comment pour-

(1) Voyez, par exemple, les courbes de distribution des salaires des ouvriers des mines de houille de Belgique, très différentes à deux époques d'observation distantes de moins de quatre ans. (*Statistique des salaires dans les mines de houille*, octobre 1896-mai 1900; publication de l'Office du Travail de Belgique, planches insérées entre les pages 34 et 35.)

rait-on dire que la courbe exprime la loi du phénomène ? Elle ne fait que représenter graphiquement et mathématiquement l'aspect du phénomène à deux époques. Le calcul mathématique définit rigoureusement les courbes en présence et à cet égard il est un moyen de représentation beaucoup plus achevé que tout autre, mais il ne peut fournir l'explication causale du phénomène. La relation qui existe entre le montant du salaire et le nombre de personnes recevant le salaire indiqué à chaque taux est simplement mathématique; la formule ne fait que préciser le contour et l'étendue de la masse considérée, elle n'a pas pour objet d'en découvrir ni d'en définir le lien causal (1).

211. *Références.*

BENINI (R.), *Principii di statistica metodologica.* Torino, 1906, pp. 227-236.

BOSCO (A.), *Lezioni di statistica.* Roma, 1909, pp. 329-357.

DAVENPORT (C.-B.), *Statistical methods with special reference to biological variation.* New York, 1904, ch. II « on the seriation and plotting of data and the frequency polygon ».

ELDERTON (W. Palin et Ethel M.), *Primer of statistics.* London, 1912, ch. III : Frequency distributions.

KING (Wilford), *The elements of statistical method.* New York, 1912, ch. XI : Frequency tables and graphs.

PEARSON (Karl), *Skew variation in homogeneous material.* Philosophical Transactions of the Royal society of London. Série A., vol. CLXXXVI (1895), pp. 343-414.

ID. *Cloudiness,* note on a novel case of frequency. Proceedings of the Royal Society, vol. LXII (1897), pp. 287-290.

ID. Supplement to a *Memoir on skew variation.* Phil. Trans. Royal Soc., série A., vol. CXCVII (1901), pp. 443-459.

SECRIST (H.), *An Introduction to statistical methods.* New York, Macmillan, 1917, ch. V, *passim.*

YULE (G. Udny), *An introduction to the theory of statistics.* London, 1911, ch. VI : the frequency distribution, pp. 83-105.

ZIZEK, *Statistical averages.* New York, 1913, ch. V : formation of magnitude, classes et III[e] partie, ch. II : the dispersion of series, etc., B., pp. 270-291.

---

(1) ZIZEK, *loc. cit.*, pp. 352-353.

CHAPITRE II

Moyennes, médiane, dominante

I. — Définitions, classification, espèces de moyennes.

212. Entre les manifestations changeantes et complexes des phénomènes collectifs, l'observateur est invinciblement amené à chercher une donnée numérique représentative de la masse. Ce besoin de simplification est universel, et répond à une nécessité de notre nature; l'esprit humain est incapable de concevoir, avec une précision suffisante, un ensemble de faits présentant une certaine complexité : les différences s'atténuent, les ressemblances s'effacent, il ne reste plus, après un certain temps, qu'une image confuse, presque impossible à définir. Veut-on cependant comparer l'une à l'autre deux observations portant sur des phénomènes collectifs, il est impossible de se contenter de notions vagues et fuyantes. Les séries doivent se caractériser, se synthétiser en une expression numérique simple. C'est à ce besoin que répondent les moyennes et leurs variantes, la médiane et la dominante (1).

M. Bowley, après avoir passé en revue les diverses moyennes, résume ainsi leur fonction générale : « Elle consiste, dit-il, à exprimer un groupe complexe au moyen de

---

(1) La dominante est l'expression proposée par M. L. March pour correspondre, en français, au terme « mode » employé par les auteurs anglais. Nous adoptons cette terminologie.

quelques nombres simples. L'esprit humain ne peut à la fois embrasser des grandeurs de millions de termes; elles doivent être groupées, classées, mises en moyennes (1). »

M. Yule, en étudiant les moyennes, commence par faire ressortir ce que les procédés de classification présentent d'insuffisant, pour caractériser la distribution des variables; la définition quantitative s'impose, surtout quand on désire établir des comparaisons entre les caractères correspondants de deux ou plusieurs séries. L'essence de la moyenne consiste donc dans son pouvoir de représentation (2).

L'impression qui se dégage de l'ouvrage de M. von Mayr est la même; pour cet auteur également, la fonction de la moyenne est de donner une seule expression simple qui contient en elle-même le résultat net de la série entière (3).

L'opinion de M. Benini ne diffère pas des précédentes : « La moyenne, dit-il, répond à un besoin de notre esprit, incapable de retenir une série exacte d'impressions se rapportant à diverses grandeurs numériques...; elle substitue à la série un terme unique, à la formation duquel concourent et collaborent tous les termes donnés (4). »

On peut dire que la statistique moderne est unanime sur ce point; la moyenne est une représentation abrégée de la série, c'est une sorte de procédé mnémotechnique appliqué aux données numériques.

213. Que la moyenne soit bien réellement un terme qui synthétise et représente tous les autres, nul n'y contredira, mais n'est-elle que cela? Et à se borner à ensisager cette

---

(1) Bowley (A. L.), *Elements of statistics*, p. 130.

(2) Yule (G. U.), *An introduction to the theory of statistics*, p. 107.

(3) Mayr (G. von), *Theoretische statistik*, p. 98.

(4) Benini (R.), *Principii di statistica*, p. 91.

face de la question, ne risque-t-on pas de simplifier, un peu plus qu'il ne conviendrait, le problème qui nous occupe?

Certains statisticiens ont poussé plus avant l'analyse. Frappés des propriétés mathématiques de la moyenne, propriétés qui étaient déjà connues des anciens, ils s'arrêtent volontiers à cet aspect des choses et en tiennent compte dans leurs définitions. L'expérience nous apprend que les résultats des mesures prises sur une quantité continue sont seulement d'une exactitude approximative. Entre les diverses mesures prises d'un même objet, la moyenne est la plus exacte et l'erreur probable de la moyenne d'un grand nombre de termes est moindre que celle de chaque terme en particulier. Ne suit-il pas de ces théorèmes, que la moyenne est quelque chose de plus que la simple représentation synthétique d'une série? Il nous semble que Lexis a parfaitement exposé ce point de vue. « Une formule purement empirique, dit-il, ne nous apprend rien de nouveau; ce n'est qu'une condensation plus ou moins commode des chiffres observés; mais, quand on réussit à démontrer que, dans un grand nombre de chiffres observés, il y en a un ou plusieurs autour desquels les autres se groupent d'après une formule établie *à priori,* on aura fait, selon moi, un pas au delà du simple empirisme des courbes paraboliques. On n'aura pas établi une loi, mais on aura simplifié la complexité des phénomènes, en les soumettant à un point de vue rationnel et à une hypothèse plausible (1). » Et plus loin : « Toutes les fois que les influences positives et négatives sont d'une possibilité égale, qu'elles sont très nombreuses et que leur intensité moyenne, mesurée par leur effet sur le rapport variable reste égale, les rapports observés, pris en assez grand nombre, se rangeront autour

---

(1) Lexis (W.) : « Sur les moyennes normales appliquées aux mouvements de la population et sur la vie normale » (*Annales de démographie internationale,* t. V, 1880, p. 481).

de leur moyenne, selon la formule exponentielle, et, même quand on n'en a qu'un nombre restreint, on en peut déduire leur écart probable avec une exactitude suffisante (1). » Ce caractère de la moyenne n'a rien de mystérieux ; « il ne fait qu'exprimer le résumé abrégé de certains faits et relations » (2).

Il y a lieu, à notre avis, de tenir compte de ce caractère et c'est pourquoi nous proposons de définir la moyenne : *l'expression de l'état quantitatif normal d'un phénomène déterminé.* En disant que la moyenne est une *expression* de l'état dont il s'agit, nous entendons marquer son caractère descriptif et nous rejetons, en même temps, toute idée d'après laquelle la moyenne serait quelque chose de substantiel ou marquerait l'action d'une force coercitive. En parlant de l'état *quantitatif,* nous limitons aux relations numériques le domaine de la moyenne. Le terme *normal* a aussi une grande signification dans la définition : il tient compte de l'action de la distribution exponentielle des variables rangés autour de la moyenne.

214. On peut, dès à présent, envisager les quantités à réunir dans les moyennes (3).

1° Une moyenne n'est pas et ne peut pas être le résultat d'une approximation. Elle suppose toujours un calcul réalisé dans de rigoureuses conditions d'exactitude. Par conséquent, les évaluations vulgaires, confondues à tort avec les moyennes, n'en sont pas, en réalité. Lorsqu'on dit : « un ouvrier menuisier gagne, en moyenne, 3 fr. 50, dans telle localité », on exprime un jugement qui, d'après la sagacité de l'observateur, a plus ou moins de chances de se

---

(1) LEXIS (W.) : « Sur les moyennes normales appliquées aux mouvements de la population et sur la vie normale » (*Annales de démographie internationale,* t. V, 1880, p. 488).

(2) Id., *loc. cit.*, p. 488.

(3) Cfr. sur ce point YULE (G. U.), *An introduction to the theory of statistics,* p. 108, et BOWLEY (A.), *Elements of statistics,* p. 130.

rapprocher de la réalité, mais qui, en l'absence des calculs nécessaires, n'a pas le caractère d'une moyenne. C'est pour cette raison que beaucoup de relevés des salaires, dans des enquêtes anciennes, sont sans valeur; ces évaluations sont élevées ou réduites d'après le degré d'optimisme ou de pessimisme du témoin, ou selon son intérêt personnel. (Cfr. les curieuses divergences entre les témoignages de patrons et d'ouvriers dans les enquêtes industrielles, par exemple dans l'enquête de la Commission du travail, en Belgique, en 1886.)

2° Elle doit être basée sur toutes les observations faites. Dans le cas contraire, elle n'est pas, en réalité, caractéristique de la distribution véritable. Lorsqu'elle est complète, elle montre quel est le type, s'il en existe un. Si l'on a pu contester, dans des cas spéciaux, le caractère typique de la moyenne, c'est que l'on n'a pas toujours eu recours à la forme de moyenne la mieux appropriée à la recherche. Beaucoup de chercheurs se sont bornés à l'emploi de la moyenne arithmétique, comme si les moyennes géométrique ou harmonique, la médiane et la dominante (*mode* des Anglais), n'existaient pas. On trouvera plus loin (Cfr. n° 264) la démonstration que la médiane, par exemple, convient mieux que la moyenne arithmétique à l'étude de la répartition des revenus et des salaires. Le simple bon sens indique que le mode de représentation doit être choisi d'après les exigences particulières du sujet à traiter.

3° Elle sera représentative, c'est-à-dire qu'elle éveillera dans l'esprit du chercheur une idée générale se rapportant à la série entière. Si l'expression de la moyenne est purement mathématique, elle ne remplira pas entièrement l'office qu'on en attend et il faudra donner la préférence à une formule de représentation peut-être moins rigoureusement exacte, mais parlant davantage à l'imagination.

4° La stabilité est aussi une de ses qualités. La moyenne ne peut pas subir de modifications si l'on introduit, dans

la série d'où elle est tirée, de légères altérations, soit dans les chiffres, soit dans les signes.

5° Il faut aussi qu'elle ne présente pas de difficultés spéciales de calcul. C'est pour cette raison qu'on est amené fréquemment à donner la préférence à la moyenne arithmétique. Il y a parfois dans ce choix une part d'exclusivisme non justifiable. Aucun calcul de moyennes ne présente de réelles difficultés de calcul et il ne serait pas admissible d'arguer de prétendues difficultés pour se limiter à l'emploi d'une formule invariable; le genre de moyenne doit être choisi d'après des considérations propres à l'objet envisagé et au but à atteindre.

215. La notion de la moyenne est double. Au point de vue mathématique, la moyenne est la valeur la plus probable entre toutes celles d'une série de données qu'on peut ou qu'on doit supposer entachées d'une erreur quelconque, même légère. On peut aussi dire, en envisageant plus spécialement le point de vue des probabilités, que la moyenne converge, à mesure que le nombre d'éléments à comparer augmente, vers une valeur spéciale et fixe, dépendant de la loi de probabilité particulière à chaque espèce de grandeur (1). Si elle n'exprime pas encore la vérité absolue, que nous sommes fréquemment impuissants à déterminer, elle en est au moins l'expression la plus complète, telle que nous pouvons la concevoir, étant donné l'imperfection de nos sens ou de nos moyens d'investigation.

La moyenne joue aussi en statistique le rôle de la synthèse, l'intelligence humaine étant, comme on l'a dit, naturellement portée à unifier en une synthèse plus ou moins compréhensive les impressions quantitatives diverses qu'elle peut recevoir des objets. Les séries forment déjà

---

(1) Cfr. Boudin-Mansion, *loc. cit.*, p. 144.

un ensemble où viennent se confondre de nombreuses données. La moyenne est une sorte de commune mesure, une expression idéale qui résume tous les termes du problème. La fonction des moyennes est d'exprimer un ensemble complexe de phénomènes au moyen d'un petit nombre de chiffres simples.

216. Au double point de vue qui vient d'être résumé correspondent deux sortes de moyennes : la moyenne *objective* ou *réelle* et la moyenne *subjective* ou *idéale*. On a proposé également de donner à la première le nom de moyenne *typique* et à la seconde celui de moyenne *indice*, mais cette distinction n'a pas été conservée par la science ; l'expression « moyenne typique » est employée par l'école contemporaine pour désigner l'expression numérique qui marque la dispersion de la série (Cfr. ch. III) et il convient d'éviter toute confusion entre cette mesure et l'expression synthétique de la série, qui est la moyenne à proprement parler.

La distinction entre la moyenne objective et la moyenne subjective n'est pas non plus universellement adoptée. L'école contemporaine anglaise, notamment, en fait assez bon marché (1). Cependant, elle répond à la nature des choses et mérite de retenir quelques instants notre attention (2).

La moyenne objective (ou réelle) est celle qui résulte des différentes mesures d'un même objet. La série synthétisée par la moyenne objective est formée de grandeurs à peu près les mêmes, se rapportant toutes à un objet unique, et dont les variations ne dépendent que de faibles erreurs commises au cours de l'observation ou de modifications lé-

---

(1) On n'en trouve trace, notamment, ni dans Venn, ni dans Bowley, ni dans Yule; Edgeworth, dans son article « Average » du dictionnaire de Palgrave, n'en fait pas non plus mention. Par contre, Stanley Jevons (*Principles of Science*) l'adopte sans contestation.

(2) On trouvera une excellente analyse de la notion des moyennes objectives et subjectives dans Zizek : *Statistical Averages*, pp. 10-13.

gères survenues pendant l'observation elle-même. Un cas classique est tiré des calculs astronomiques ayant pour objet, par exemple, la détermination de l'ascension droite d'un astre. On pourrait supposer qu'une observation aussi usuelle ne prête guère à erreur; cependant, Quetelet qui, en même temps que statisticien, était astronome, a pris soin de nous avertir des causes nombreuses d'erreur qui existent dans l'astronomie de position : « quelque précis que soit l'instrument, dit-il, il n'est point parfait dans toutes ses parties; quelles que soient l'adresse et l'expérience de l'observateur, son coup d'œil n'est pas infaillible; l'air peut être dans des circonstances plus ou moins défavorables; nous ne voyons les astres que du fond de l'atmosphère dans laquelle nous sommes plongés et, à cause des réfractions, ils ne sont réellement pas dans les lieux où nous les apercevons (1). » Aussi les déterminations de l'ascension droite de la polaire faites, au nombre de 487, à l'observatoire royal de Greenwich, pendant les années 1836 à 1839 inclusivement, n'ont-elles pas toutes donné des résultats identiques. En combinant ces résultats, on détermine une moyenne objective résultant des différentes mesures d'un même objet.

Les mesures dont on tire la moyenne objective doivent, comme dit M. Mansion, leur inégalité à leur inexactitude. Ces mesures expérimentales sont nécessairement liées entre elles par une loi de continuité telle qu'on doit admettre que, s'il n'y a pas de cause permanente d'erreur, il y a autant de chances de voir se produire des erreurs en plus (positives) et des erreurs en moins (négatives). La courbe de probabilité prend, dans ce cas, une forme symétrique par rapport à l'ordonnée centrale représentant la valeur exacte, ou, ce qui est la même chose, une erreur nulle (2).

---

(1) QUETELET, *Lettres sur la théorie des probabilités*, p. 124.

SION, *loc. cit.*, p. 148.

217. A cette moyenne s'oppose la moyenne subjective (ou idéale). Elle vient des diverses mesures se rapportant, non à un même objet, mais à plusieurs objets homogènes. C'est la véritable moyenne usitée en statistique, particulièrement dans la statistique sociale. On n'a presque jamais à rechercher quelle est la qualité réellement exacte parmi une série de grandeurs observées, mais on a très fréquemment à caractériser par une expression simple une série plus ou moins étendue de mesures se rapportant à des objets multiples, mais homogènes. Les statistiques de salaires fournissent un exemple convenable de ce genre de recherches. On demande, par exemple, quel est le salaire moyen d'ouvriers mâles adultes travaillant dans une industrie déterminée, à une date fixée et dans un endroit indiqué. Les objets qui sont à caractériser ici sont les salaires effectivement gagnés; il y a autant d'observations qu'il y a de salaires, et ces observations seront homogènes si l'on a pris soin de décider qu'il s'agit d'une certaine catégorie d'ouvriers, observés au même moment, dans un même lieu et dans une industrie nettement déterminée. Cette quantité conventionnelle sera représentative de l'ensemble, comme la moyenne objective est la synthèse des diverses observations relatives à l'ascension droite de l'astre observé.

218. En comparant ces deux espèces de moyennes, nous voyons qu'elles diffèrent substantiellement. La moyenne objective s'applique à un seul objet qui est supposé n'avoir subi aucune modification essentielle au cours de l'observation. En conséquence, les différentes mensurations qui ont été faites ne diffèrent entre elles que par leur degré de précision. Les instruments employés n'ont peut-être pas toujours fonctionné d'une manière parfaite et les écarts autour de la moyenne expriment les erreurs dues à cette cause; ou bien, l'observateur n'a pas apporté un soin égal dans toutes ses vérifications, et alors les écarts indiquent un coefficient d'erreur personnelle; ou, ce qui est le cas le plus

fréquent, les deux causes d'erreur coexistent et se combinent avec une troisième ou une quatrième. La moyenne objective a pour but d'indiquer quelle est la dimension véritable de l'objet de l'étude.

Il ne peut être question de cela dans la moyenne subjective. Les objets homogènes qui la composent sont tous différents ; leurs dimensions sont supposées connues avec toute l'exactitude désirable, mais on voudrait obtenir, par le calcul, une mesure commune qui les synthétise, de façon à remplacer l'énumération de la série par une expression abrégée. Cette expression est simplement idéale (de là le second nom sous lequel on désigne la moyenne subjective). Elle n'exprime aucun caractère de grandeur réelle et elle n'indique pas non plus la gravité de l'erreur provenant soit de l'observateur, soit de l'instrument d'observation ; sa portée est donc essentiellement synthétique et représentative (1).

Les deux espèces de moyennes diffèrent encore sous le rapport de la courbe applicable à la série. La moyenne objective dérive d'une série dont le développement suit de très près la courbe normale des erreurs. Les termes peuvent être considérés comme des valeurs empiriques entachées d'erreurs plus ou moins grossières. Dans cette conception, on se rapproche de la théorie de Gauss, dont nous avons déjà dit quelques mots plus haut. (Cfr. n° 202.) La répartition des erreurs est supposée se faire, d'après la théorie mathématique, conformément à l'expérience et à la raison, c'est-à-dire que les erreurs les plus graves sont les moins nombreuses, tandis que le nombre de celles ayant peu d'im-

---

(1) M. Mansion, *loc. cit.*, p. 149, écrit à ce propos : « On peut faire avantageusement usage des moyennes arithmétiques et des limites entre lesquelles sont comprises les quantités dont elles dérivent, toutes les fois que l'on veut restreindre le vague d'une expression collective et lui donner de l'individualité; mais il importe de ne pas perdre de vue qu'il ne s'agit que d'une simple opération de calcul entre des quantités qui n'ont entre elles aucune relation essentielle. »

portance augmente progressivement. En général, on peut dire que la distribution des termes formant la moyenne objective se fait d'après la formule de distribution symétrique. Au contraire, les séries d'observations se rapportant à des objets différents marquent une distribution asymétrique autour de la moyenne.

219. Nous avons défini plus haut (Cfr. n° 207) le caractère des courbes asymétriques rappelant plus ou moins la courbe de Gauss, nous n'y reviendrons plus ici. Nous nous bornerons à signaler que les séries de mesures se rapportant à plusieurs objets homogènes appartiennent à cette espèce de courbes. On les rencontre en très grand nombre dans le domaine des sciences morales et politiques; les exemples appartenant à cette série de questions sont pour ainsi dire inépuisables (1). On sait peut-être moins que les sciences naturelles en ont fourni aussi de nombreux spécimens. Au cours de travaux extrêmement intéressants, qui se sont multipliés ces dernières années, des naturalistes et des botanistes ont prouvé par des recherches expérimentales, conduites avec une grande rigueur scientifique, que les lois de variation autour de la moyenne se retrouvaient, dans une certaine mesure, appliquées dans la sphère des phénomènes biologiques et botaniques (2). Voici, pour 249 exemplaires du *Pimpinella Saxifraga* (variété : feuilles à grandes folioles), observés par nous à Berg-lez-Tongres (Limbourg belge) au mois de septembre 1916 (3), la répar-

(1) Nous en avons donné quelques spécimens, accompagnés de graphiques, au chapitre précédent.

(2) On en trouvera une bibliographie très étendue dans l'ouvrage de C. B. Davenport déjà cité : *Statistical methods*, pp. 85-104.

(3) Nous sommes redevables de la détermination de cette plante à l'obligeance de M. E. De Wildeman, directeur du Jardin botanique de l'Etat, à Bruxelles, à qui nous sommes heureux d'adresser nos remerciements.

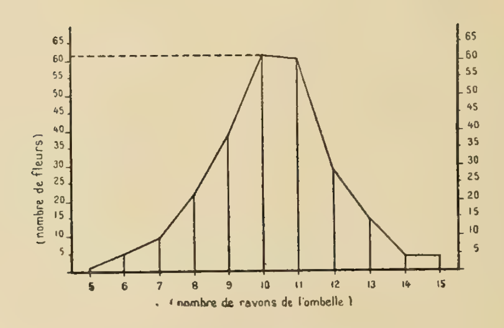
tition des ombelles comptant le nombre de rayons indiqués ci-dessous (1) :

EXEMPLE 1.

Pimpinella Saxifraga L.

(A. JULIN, Berg, Septembre 1916.)

Nombre de rayons	Nombre d'exemplaires possédant le nombre de rayons ci-contre	Nombre de rayons	Nombre d'exemplaires possédant le nombre de rayons ci-contre
5	1	11	61
6	5	12	29
7	9	13	14
8	22	14	4
9	38	15	4
10	62		249



Pimpinella Saxifraga L.

FIG. 21.

La longueur des graines de caféier donne aussi naissance à une courbe asymétrique dérivée de la courbe de Gauss.

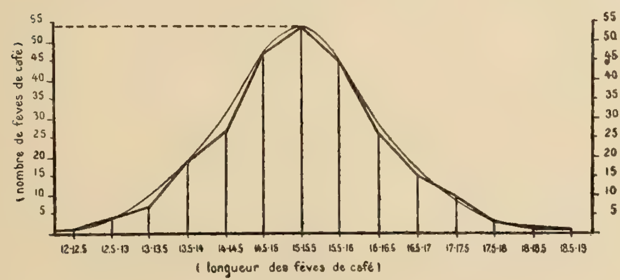
(1) Cfr. les expériences portant sur le *Pimpinella Saxifraga*, de Ludwig : « Ueber Variationskurven und Variationsflüche der Pflanzen. Botanisch-statistische Untersuchungen. (*Botanisch Centralblatt*, Band. LXIV, n° 3, pp. 65-67.)

Voici les chiffres obtenus par le Dr Cramer, d'après les graines recueillies sur le *Liberia Coffea* (1) :

Exemple 2. — Liberia Coffea.

(Dr Cramer, Indes néerlandaises.)

Longueur des fèves	Nombre d'exemplaires ayant les dimensions ci-contre	Somme des longueurs moyennes
12 — 12.5 millimètres . . .	1	12.25
12 5 — 13 » . . .	4	51 00
13 — 13.5 » . . .	7	92.75
13.5 — 14 » . . .	19	261.25
14 — 14.5 » . . .	27	384.75
14.5 — 15 » . . .	47	693 25
15 — 15.5 » . . .	54	823.50
15.5 — 16 » . . .	45	708.75
16 — 16.5 » . . .	26	422.50
16.5 — 17 » . . .	15	251.25
17 — 17.5 » . . .	10	172.50
17.5 — 18 » . . .	3	53.25
18 — 18.5 » . . .	1	18.25
18.5 — 19 » . . .	1	18.75
	260	3,964.00



Liberia Coffea.

Fig. 22.

---

(1) Dr. P. I. S. Cramer, Mededeelingen uitgaande van het Departement van Landbouw, n° 11. Gegevens over de variabiliteit van de in Nederlandsch-Indië verbouwde koffiesoorten. Kolff & C°, Batavia, 1913, p. 40.

Ludwig, dans son travail sur les courbes et surfaces de variation des plantes, cité plus haut (cfr. p. 358, note 1), a fait de très nombreuses expériences sur une série de végétaux; les courbes qu'il a tracées présentent parfois plusieurs « modes » (1); c'est le cas, notamment, pour le *Pimpinella saxifraga L.*, que l'auteur a analysé, mais la réunion de certains résultats fragmentaires permet la construction d'une vraie courbe asymétrique. Ludwig a trouvé pour cette plante des courbes comprenant plusieurs modes, mais en les réunissant il en a tracé une nouvelle où les deux sommets primitifs existaient encore, mais où, en même temps, la dépression intermédiaire est remplacée par un large plateau aux points 10 et 11. Ce résultat est d'accord avec celui auquel nous sommes arrivé (2).

220. La classification des moyennes peut aussi se faire d'après la façon de les calculer. C'est même la classification la plus connue du grand public; les traités d'algèbre et la pratique ont rendu familières à chacun les notions de : moyenne arithmétique, moyenne géométrique, moyenne harmonique. On connaît moins la moyenne anti-harmonique et les propriétés réciproques que montrent les moyennes les unes à l'égard des autres. Bien que les moyennes dont nous faisons usage soient très peu nombreuses, nous savons qu'en réalité il n'y a pas de limites dans cette direction; on peut, en effet, imaginer des valeurs moyennes de formes diverses et en nombre considérable (3).

On fait remonter à Pythagore l'introduction des trois moyennes classiques : l'arithmétique, la géométrique et l'harmonique. Boëtius en avait proposé trois autres qui étaient le contraire de la géométrique et de l'harmonique,

---

(1) Sur le sens de cette expression, cfr. p. 333, note 2.

(2) Ludwig, *loc. cit.*, et de Bruycker, *De statistische methode in de plantkunde*. Gent, Siffer, 1910, p. 57.

(3) Messedaglia, « Calcul des valeurs moyennes » (*Annales de démographie internationale*, 4e année (1880), p. 388).

et puis, finalement, quatre nouvelles. Le mathématicien Jordanus (fin du XIIe siècle) en ajouta même une onzième (1).

L'accord s'est fait pour ne pas encombrer inutilement la pratique de formules trop compliquées ou ne présentant qu'un simple intérêt théorique. Aujourd'hui, les statisticiens se bornent, en général, à utiliser les moyennes classiques; leur importance, comme leur valeur pratique, peut être exprimée dans cet ordre : en tête, la moyenne arithmétique; en second lieu, la moyenne géométrique; enfin, la moyenne harmonique; la moyenne contre-harmonique et la moyenne quadratique jouent un rôle spécial. Aux moyennes classiques on a ajouté, à une époque relativement récente, deux expressions synthétiques nouvelles : la médiane et la dominante, dont nous parlerons plus loin. (Cfr. même chapitre, VI et VII.)

221. La *moyenne arithmétique* s'écrit au moyen de formules diverses, dont voici les principales (2) :

$$M = \frac{a_1 + a_2 + a_3 + \dots + a_n}{n} \qquad (8)$$

formule dans laquelle M = la moyenne; $a_1, a_2, a_3 \dots a_n$ les termes de la série de laquelle dérive la moyenne; n, le nombre de termes de la série.

On peut aussi écrire :

$$M = \frac{1}{N}(X_1 + X_2 + X_3 + \dots + X_n) \qquad (9)$$

formule où l'on peut remplacer la succession des X par le symbole Σ qui désigne la somme des quantités de la série, de sorte que l'on a :

$$M = \frac{1}{N}\Sigma(X) \qquad (10)$$

(1) STANLEY JEVONS, *The Principles of Science.* London, 1892, p. 360.

(2) Nous employerons la lettre M pour désigner la moyenne; les Anglais et les Américains font souvent usage de la lettre A (average). Les chiffres entre parenthèses qui suivent la formule indiquent le n° d'ordre de celle-ci.

Lorsque la moyenne est calculée sur une série composée de classes différentes ayant chacune leur fréquence propre (par exemple, des taux de salaires gagnés par des ouvriers en nombre variable), la formule de la moyenne est :

$$M = \frac{1}{N} \Sigma \; (f . x) \qquad (11)$$

où f signifie la fréquence et x la valeur de la classe, pratiquement la demi-valeur des deux limites extrêmes (1.25 fr. à 1.50 fr. = 1.375 fr.).

La *moyenne géométrique* est également d'un usage courant, quoiqu'elle soit moins fréquemment employée que la moyenne arithmétique. Nous la désignons par la lettre G.

On l'écrit :

$$G = \sqrt[n]{a_1 \,.\, a_2 \ldots a_n} \qquad (12)$$

ou encore :

$$G = (X_1 \,.\, X_2 \,.\, X_3 \ldots X_n)^{\frac{1}{n}} \qquad (13)$$

La recherche s'effectuant au moyen des logarithmes, on peut aussi avoir :

$$\log. G = \frac{1}{N} \Sigma \; (\log. X) \qquad (14)$$

d'où il résulte que le logarithme de la moyenne géométrique d'une série de valeurs est la moyenne arithmétique de leurs logarithmes.

La *moyenne harmonique* est d'un emploi peu fréquent. Nous la désignons au moyen de la lettre H.

$$H = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}} \qquad (15)$$

Ou encore, quand on envisage le cas de deux valeurs seulement :

$$H = \frac{2(a_1 + a_2)}{a_1 + a_2} \qquad (16)$$

formule identique à (15).

La *moyenne contre-harmonique* (CH) a pour formule générale :

$$CH = \frac{a_1^2 + a_2^2 + a_3^2 + \dots + a_n^2}{a_1 + a_2 + a_3 + \dots + a_n} \qquad (17)$$

La *moyenne quadratique* (M^2) s'exprime par la formule :

$$M^2 = \left(\sqrt{\frac{a_1^2 + a_2^2 + a_3^2 + \dots a_n^2}{n}}\right) \qquad (18)$$

222. Avant d'exposer les caractères de chacune de ces moyennes, nous ferons une brève application du calcul aux nombres : 34, 38, 42, 46, afin de familiariser le lecteur avec l'emploi des formules. En regard du calcul se trouve, entre parenthèses, le numéro de la formule employée.

$$M = \frac{34 + 38 + 42 + 46}{4} = \frac{160}{4} = 40 \qquad (8)$$

Admettons que ces données soient affectées des poids : 4, 3, 2, 1.5.

Nous avons :

$$M = \frac{34 \times 4 + 38 \times 3 + 42 \times 2 + 46 \times 1.5}{10.5} = \frac{136 + 114 + 84 + 69}{10.5}$$

$$= \frac{403}{10.5} = 38{,}381 \qquad (10)$$

La moyenne géométrique de ces nombres s'exprime par la

moyenne de leurs logarithmes ramenée à son nombre naturel :

$$G = \frac{1.53147892 + 1.57978360 + 1.62324929 + 1,66275783}{4}$$

$$= \frac{6.39726964}{4} = 1.59931741 = 39,750 \qquad (14)$$

La moyenne harmonique des mêmes chiffres est :

$$H = \frac{1}{\frac{0.029411765 + 0.026315789 + 0.023809524 + 0.021739130}{4}}$$

$$= \frac{1}{0.025319052} = 39,498 \qquad (15)$$

Leur moyenne contre-harmonique est :

$$C.H = \frac{1156 + 1444 + 1764 + 2116}{34 + 38 + 42 + 46} = \frac{6480}{160} = 40,5 \qquad (17)$$

Enfin, leur moyenne quadratique est :

$$M^2 = \sqrt{\frac{1156 + 1444 + 1764 + 2116}{4}} = \sqrt{\frac{6480}{4}}$$

$$= \sqrt{1620} = 40,2492236 \qquad (18)$$

Indépendamment de ces moyennes, il y a lieu de mentionner la médiane et la dominante (ou mode), qui sont des moyennes de position. Ces deux moyennes sont d'un usage récent. Elles ont été mises en évidence principalement par l'école statistique anglaise. Un paragraphe spécial est réservé à chacune d'elles à la fin de ce chapitre.

223. La classification des moyennes d'après le mode de calcul qui leur est applicable est adoptée généralement. Cependant, il y a lieu de signaler un classement différent proposé par Fechner (1). Cet auteur divisait les moyennes en trois groupes :

---

(1) Fechner, *Kollektivmasslehre* et *Ueber den Ausgangswert der Kleinsten Abweichungssumme*, pp. 75 et 37.

1° Celles constituées par une valeur telle que la somme absolue des carrés des écarts à la moyenne est un minimum par rapport à celle-ci. La moyenne arithmétique appartient à ce type ;

2° Le groupe de moyennes constitué par celles de la forme

$$Mn = \sqrt[n]{\frac{\Sigma a^n}{m}} \tag{19}$$

équation dans laquelle n indique la puissance, Σ la somme des termes de la série, a les termes, m le nombre de ces termes ; la moyenne quadratique appartient au second ordre de ce groupe

3° Enfin, les moyennes de combinaison.

Cette classification, d'essence mathématique plutôt que statistique, n'a pas détrôné la première.

II. — La moyenne arithmétique.

224. La moyenne arithmétique est à la fois la plus usitée et, dans de nombreux cas, la plus sûre de toutes les expressions synthétiques employées pour caractériser une série.

On la divise en moyenne arithmétique simple et en moyenne arithmétique composée ou pondérée. En réalité, il n'y a pas de distinction essentielle entre ces deux formes ; la règle appliquée est la même dans les deux cas, elle ne diffère que par une simple modalité. Dans la moyenne arithmétique simple, chaque terme peut être considéré en lui-même, tandis que dans la moyenne pondérée, il y a lieu d'observer leur fréquence, ce qui oblige nécessairement à tenir compte des « poids » dans le calcul. Ainsi, dans une brasserie, on produit X hectolitres de bière par an ; la moyenne de la production par semaine (si l'on a travaillé chaque semaine) sera $\frac{X}{52}$, moyenne arithmétique simple ; pour la valeur moyenne de l'hectolitre, on sera obligé de

tenir compte des quantités de bière qui ont été vendues à différents prix : soit X_1, X_2, X_3 ces quantités et v_1, v_2, v_3 les prix, on a :

$$\frac{X_1 \times v_1 + X_2 \times v_2 + X_3 \times v_3}{X_1 + X_2 + X_3}, \qquad (11)$$

moyenne arithmétique pondérée.

La règle de la moyenne arithmétique est fondée sur le principe de compensation équitable. « Dans les faits qu'étudie la statistique, a écrit M. Lucien March dans un remarquable travail (1), toute vue générale implique l'hypothèse que des changements survenus dans les circonstances nombreuses qui gouvernent les phénomènes, la plupart sont assez faibles ou se compensent suffisamment pour ne point altérer sensiblement le résultat d'une observation éclairée. » Ainsi, d'après le même auteur, « le principe de compensation sur lequel est basée toute comparaison de faits collectifs, n'est que la généralisation de la formule d'un marché équitable que les hommes ont sans doute adoptée depuis qu'ils vivent en société. La règle de la moyenne qui, du point de vue logique, est fondée sur une extension du principe de raison suffisante traduit cette formule (2). »

Il en résulte que « la moyenne arithmétique ou, plus brièvement, la moyenne de n quantités est la grandeur qui, répétée n fois, fournit le total de n quantités » (3).

225. La moyenne arithmétique s'obtient en divisant la somme des termes par le nombre des termes. Rappelons son expression générale :

$$M = \frac{a_1 + a_2 + a_3 + \dots + a_n}{n} \qquad (8)$$

Elle est d'un usage extrêmement fréquent.

---

(1) Lucien March : « Essai sur un mode d'exposer les principaux éléments de la théorie statistique ». (*Journal de la Société de statistique de Paris*, 51e année, 1910, pp. 447-486.)

(2) Lucien March, *loc. cit.*, p. 448.

(3) Id., *loc. cit.*, p. 449.

Les observations hebdomadaires faites à l'aide des thermomètres, dans les observatoires, sont, par exemple, réduites en moyenne mensuelle. A l'observatoire d'Edimbourg, les lectures faites à l'un des thermomètres (t) ont donné, en janvier 1899, les résultats suivants : 47.50 — 47.49 — 47.43 — 47.36 — 47.32 (1). D'où la moyenne mensuelle pour janvier 1899 :

$$Mt_{,} = \frac{47.50 + 47.49 + 47.43 + 47.36 + 47.32}{5} = 47.42$$

Les moyennes sont nécessaires en démographie pour apprécier les variations qui, d'une période à l'autre, s'observent dans les phénomènes. Voici, par exemple, le nombre de mariages contractés en Angleterre et dans le pays de Galles, pendant la période décennale 1891-1900 (2) :

Exemple 3. — **Mariages en Angleterre et dans le pays de Galles (1891-1900).**

Années.	Nombre de mariages.	Années.	Nombre de mariages.
1891	226,526	1896	242,764
1892	227,135	1897	249,145
1893	218,689	1898	255,379
1894	226,449	1899	262,334
1895	228,204	1900	257,480

D'après la formule :

$$M = \frac{1}{N} \Sigma (X) \qquad (10)$$

nous avons :

$$M = \frac{2394105}{10} = 239,410.5$$

(1) « Observations of the Edinburgh Rock Thermometers », by Thomas Heath, B. A., Assistant Astronomer. (*Transactions of the Royal Society of Edinburgh*, t. 40, 1900-1901, p. 169.)

(2) *Statistique internationale du mouvement de la population*. Paris, Imprimerie Nationale, 1907, pp. 2-3.

226. La formule de la moyenne composée ou pondérée ne diffère pas de celle de la moyenne arithmétique simple, sauf qu'ici il faut nécessairement tenir compte de l'importance respective des termes. La moyenne pondérée est très fréquente dans les recherches relevant de l'ordre social ou économique.

Un cas classique de l'emploi de cette moyenne se trouve dans la formation des *index-numbers* de prix (1). Si l'on veut évaluer les variations générales de l'ensemble des prix, à une époque donnée, il convient non seulement de réunir des renseignements sûrs concernant un grand nombre d'articles, mais il faut, de plus, attribuer à chaque article un « poids » en rapport avec l'importance qu'il a dans l'ensemble des marchandises considérées. Pour obtenir ce résultat, on donne à chaque denrée une cote et on multiplie les prix par cette cote.

Les *Index-Numbers* dressés par la Commission des Finances du Sénat des Etats-Unis d'Amérique ont été calculés d'après trois méthodes dont deux sont basées sur la moyenne arithmétique pondérée. Les « poids » sont proportionnés à l'importance de la consommation parmi 2,561 familles normales dont les dépenses ont été relevées au moyen de budgets. Par rapport à 10,000, les dépenses représentent, par catégories d'articles, les chiffres proportionnels ci-après : loyer : 1,506; nourriture : 4,103; chauffage : 500; vêtement : 1,531; éclairage : 90; autres dépenses : 2,270 (2).

---

(1) On appelle « Index-Numbers », un système dans lequel les prix d'un certain nombre de marchandises considérées comme représentatives de l'ensemble des transactions sont relevés et représentés par l'unité; une année ou une moyenne d'années, ayant été prise comme base de comparaison, les variations des prix les années suivantes s'expriment par pourcentages calculés sur les prix de la première année.

(2) U. S. Department of Labor. Bureau of Labor Statistics. Numbers of wholesale prices in the United States and foreign countries, July 1910. Washington.

Par rapport à l'année de base (1860 = 100) les chiffres de l'Index en 1891 donnent les résultats consignés à la seconde colonne. La moyenne proportionnée ou pondérée se calcule donc comme suit :

Exemple 4. — **Période de base : 1860 = 100.**

GROUPES	Importance	Nombre-Index en 1891	Résultats
Loyer	1,506	100,0	1,506,000
Nourriture	4,103	103.7	4,254,811
Chauffage	500	98,1	490,500
Eclairage	90	48,1	43,290
Vêtement	1,531	75,1	1,162 029
Autres dépenses . . .	2,270	95,3	2 164,096
Totaux. . .	10,000	96,2	9,620,726

Représentant par p_0 les prix une année prise comme base, par p_x les prix les autres années, par q les quantités physiques admises comme poids, et établissant les relations entre les prix et les quantités chaque année par rapport à l'année de base, on a :

$$\frac{p_x \cdot q + p'_x \cdot q' + \ldots\ldots + p_x^n \cdot q^n}{p_0 \cdot q + p'_0 \cdot q' + \ldots\ldots + p_0^n \cdot q^n}$$

227. Dans l'exemple précédent, la moyenne est multipliée par un nombre entier exprimé par un seul chiffre, mais il arrive souvent que la classe des données est comprise entre deux chiffres limites ; tel est le cas qui se présente dans les statistiques de salaires où des nombres donnés d'ouvriers gagnent un salaire compris entre telle somme et telle autre. Dans ce cas, on adopte, comme représentant la classe, le chiffre intermédiaire entre les deux limites, dans l'hypothèse que la répartition des unités se fasse d'une manière égale.

Voici un exemple emprunté à la statistique des salaires dans les industries des métaux, publié par l'Office du Travail de Belgique (1) :

Exemple 5. — **BELGIQUE. - Salaires des ouvriers mâles adultes dans les ateliers de robinetterie.** (Octobre 1903.)

Taux des salaires	Taux intermédiaire	Nombre d'ouvriers	Montant des salaires
Francs.			Francs.
1.50 à 1.74	1,625	17	27.62
1.75 à 1.99	1,875	3	5.62
2.00 à 2.24	2,125	10	21.25
2.25 à 2.49	2,375	8	19.00
2 50 à 2 74	2,625	34	89.25
2.75 à 2.99	2,875	28	80.50
3.00 à 3.24	3,125	34	106.25
3.25 à 3.49	3,375	26	87.75
3.50 à 3.74	3,625	47	170.37
3.75 à 3 99	3,875	44	170.50
4.00 à 4.24	4,125	60	247.50
4.25 à 4.49	4,375	24	104 90
4.50 à 4.74	4,625	27	124.87
4.75 à 4.99	4,875	9	43.87
5.00 à 5.24	5,125	22	112.75
5.25 à 5 49	5,375	8	43.00
5.50 à 5.74	5.625	5	28.12
5 75 à 5 99	5,875	13	76.37
6.00 à 6.24	6,125	8	49.00
6.25 à 6.49	6,375	2	12.75
6.50 à 6.74	6,625	1	6.62
6 75 à 6.79	6,875	3	20.62
		433	1,648.48

(1) Royaume de Belgique. Office du travail. *Statistique des salaires dans les industries des métaux au mois d'octobre 1903.* Bruxelles, 1907, p. 647.

Le tableau précédent présente les résultats numériques de la formule :

$$M = \frac{1}{N} \Sigma (f.\, x) \qquad (11)$$

d'après laquelle :

$$\frac{1.625 \times 17 + 1,875 \times 3 + + \ldots}{433} = \frac{1,648.48}{433} = \text{fr. } 3,807.$$

228. La même formule est employée pour le calcul des caractères de variabilité en botanique. Calculons le nombre moyen d'ombellules pour les 249 exemplaires du *Pimpinella saxifraga L.* que nous avons analysés et dont nous avons reproduit plus haut (cfr. n° 219) la courbe caractéristique.

EXEMPLE 6. — **Pimpinella saxifraga L.**

Nombre moyen de rayons

Classement d'après le nombre de rayons	Nombre d'exemplaires possédant le nombre de rayons ci-contre	Nombre total de rayons
5	1	5
6	5	30
7	9	63
8	22	176
9	38	342
10	62	620
11	61	671
12	29	348
13	14	182
14	4	56
15	4	60
	249	2,553

$$\frac{5 \times 1 + 6 \times 5 + 7 \times 9 + + \ldots}{249} = \frac{2,553}{249} = 10.253$$

229. Il existe une autre méthode, pour calculer la moyenne arithmétique, qu'on peut désigner sous le nom de méthode abrégée ou de méthode indirecte. Elle est fondée sur cette propriété remarquable des moyennes d'après laquelle la somme algébrique des écarts à la moyenne est égale à zéro. (Cfr. plus loin V, propriétés mathématiques des moyennes, n° 246) (1).

Appelons A la valeur choisie d'une manière arbitraire, M la moyenne, z l'écart entre A et M. Nous pouvons écrire (2)

$$M = A + z \qquad (20)$$

Si f désigne le nombre de fréquences, x les valeurs de chaque classe, nous avons :

$$\Sigma\,(f.\,x) = \Sigma\,(f.\,A) + \Sigma\,(f.\,z)$$

et puisque A est une valeur constante :

$$M = A + \frac{1}{N}\Sigma\,(f.\,z) \qquad (21)$$

au lieu de calculer $\Sigma\,(f.\,x)$, on calcule donc $\Sigma\,(f.\,z)$.

L'avantage de la méthode consiste en ce que les fréquences doivent être multipliées seulement par des nombres peu importants, donc faciles à calculer; en prenant leur origine 0 à la classe choisie d'une manière arbitraire, ils ne s'élèvent qu'à quelques unités correspondant au nombre de classes au-dessus et au-dessous de la valeur arbitraire. Pour rendre ce nombre aussi petit que possible, on a soin de prendre comme point d'origine une valeur située vers le milieu de la série.

La règle peut être formulée comme suit : *la moyenne arithmétique simple d'une série de termes peut être trouvée en admettant qu'une valeur arbitrairement choisie est la*

---

(1) Cfr. sur ce point King (W.) : *The elements of statistical methods*, New-York, 1912, pp. 134-136, et Yule (U. G.) : *An introduction to the theory of statistics*, London, 1911, pp. 110-113.

(2) Cfr. Udny Yule, *loc. cit.*, p. 110.

moyenne exacte; calculez alors la somme algébrique des écarts de chacun des termes par rapport à la moyenne supposée; divisez cette somme par le nombre de termes et ajoutez le quotient au chiffre de la moyenne arbitrairement choisie. Le résultat donne la valeur de la moyenne exacte.

Faisons application de cette méthode à l'exemple 3. (Cfr. n° 225.)

Exemple 7. — **Nombre moyen de mariages de 1891 à 1900 en Angleterre et dans le pays de Galles.**

Années.	Termes de la série.	Moyenne arbitraire.	Écarts de chaque terme à la moyenne arbitraire.
1891	226,526		— 13,474
1892	227,135		— 12,865
1893	218,689		— 21,311
1894	226,449		— 13,551
1895 . . .	228,204	240.000	— 11,796
1896	242,764		+ 2,764
1897	249,145		+ 9,145
1898	255,379		+ 15,379
1899	262,334		+ 22,334
1900	257,480		+ 17,480
			Σ = — 5,895

$$-5{,}895 : 10 = -589.5$$

$$240{,}000 + (-589.5) = 239{,}410.5$$

Ce chiffre exprime la moyenne arithmétique exacte, telle qu'elle a été calculée plus haut.

230. Pour la moyenne arithmétique composée, la règle est différente : *choisissez une classe des variables comme renfermant, par hypothèse, la moyenne et désignez cette classe par 0; attribuez ensuite aux classes supérieures et*

inférieures un numéro d'ordre en partant de la classe 0; multipliez les fréquences par ce chiffre en attribuant au produit une valeur positive lorsque sa classe est supérieure, une valeur négative lorsque sa classe est inférieure à celle choisie par hypothèse. Faites la différence des deux produits; le reste forme le numérateur d'une fraction dont le dénominateur est le nombre de fréquences; réduisez la fraction, en tenant compte de l'intervalle de la classe et ajoutez le quotient avec son signe à la moyenne supposée : le total est la moyenne véritable.

Traitons par cette méthode les données de l'exemple 5.

Supposons, par hypothèse, que la moyenne est comprise dans la classe fr. 3.50 à 3.74. Nous disposons les calculs comme suit :

Exemple 8. — **BELGIQUE**. — **Salaires des ouvriers mâles adultes dans les ateliers de robinetterie.** (Octobre 1903.)

Classes des salaires	Fréquences (f)	Ecarts (δ)	Produits ($f.\ \delta$)
Francs.			
1,625	17	— 8	136
1,875	3	— 7	21
2,125	10	— 6	60
2,375	8	— 5	40
2,625	34	— 4	136
2,875	28	— 3	84
3,125	34	— 2	68
3,375	26	— 1	26
3,625	47	0 (A)	— 571
3,875	44	+ 1	44
4,125	60	+ 2	120
4,375	24	+ 3	72
4,625	27	+ 4	108
4,875	9	+ 5	45
5,125	22	+ 6	132
5,375	8	+ 7	56
5,625	5	+ 8	40
5,875	13	+ 9	117
6,125	8	+10	80
6,375	2	+11	22
6,625	1	+12	12
6,875	3	+13	39
	433		+ 887

$$\Sigma\,(f.\ \delta) = +\ 887 - 571 = +\ 316$$

$M - A$ (1) $= + \frac{316}{433} = + \frac{729}{1,000}$ ou $+\ 0,182$ (en tenant compte de l'intervalle (0.25) de la classe).

$$M = 3,625 + 0,182 = 3,807. \text{ (Cfr., n° 227.)}$$

(1) A est la valeur choisie arbitrairement comme représentant la moyenne.

Cette méthode étant importante, nous en faisons une nouvelle application à notre exemple 6, où la moyenne arithmétique se fixe par le procédé ordinaire à 10.253. (Cfr. n° 228.) Supposons que la moyenne se trouve dans la classe 11 :

EXEMPLE 9. — **Pimpinella Saxifraga L.**
Nombre moyen de rayons

Classes par rayons	Fréquences (f)	Ecarts (δ)	Produits ($f. \delta$)
5	1	— 6	— 6
6	5	— 5	— 25
7	9	— 4	— 36
8	22	— 3	— 66
9	38	— 2	— 76
10	62	— 1	— 62
11	61	0 (A)	—271
12	29	+ 1	29
13	14	+ 2	28
14	4	+ 3	12
15	4	+ 4	16
	249		+ 85

$$\Sigma \ (f. \ \delta) = -271 + 85 = -186$$

$$M - A = -\frac{186}{249} = -0.747$$

$$M = 11 + (-0.747) = 10.253$$

Dans la réduction de la fraction, nous n'avons pas dû transformer le quotient, l'intervalle de classe étant, dans ce cas-ci, égal à l'unité.

231. Il est clair que si la position de la moyenne hypothétique est modifiée, tous les calculs changent, mais le résultat reste le même. Ludwig (1) a trouvé des courbes de *Pimpinella saxifraga L* dont l'optimum était à 8; adoptons

(1) LUDWIG: *Ueber Variationskurven and Variationsflächen*, etc., pp. 65-67.

ce chiffre comme moyenne hypothétique. Les calculs se disposent comme suit :

Exemple 10. — Pimpinella Saxifraga L.
Nombre moyen de rayons

Classes par rayons	Fréquences (f)	Ecarts (δ)	Produit ($f.\ \delta$)
5	1	— 3	3
6	5	— 2	10
7	9	— 1	9
8	22	0	— 22
9	38	+ 1	38
10	62	+ 2	124
11	61	+ 3	183
12	29	+ 4	116
13	14	+ 5	70
14	4	+ 6	24
15	4	+ 7	28
	249		+ 583

$$\Sigma\,(f.\ \delta) = +\,583 - 22 = +\,561$$

$$M - A = +\frac{561}{249} = +\,2.253$$

$$M = 8 + 2.253 = 10.253$$

L'avantage que présente le procédé indirect est très appréciable : il dispense de calculs parfois fort longs lorsque les classes sont nombreuses et que les fréquences sont exprimées par des nombres considérables ; au lieu de procéder à des calculs écrits, on peut souvent se borner, dans le procédé indirect, à un calcul mental dont on écrit directement le résultat. Par contre, le procédé indirect ne conduit qu'à une approximation lorsque les écarts ne sont pas assez rapprochés ou lorsque la répartition des unités à l'intérieur de la classe ne s'effectue pas d'une manière uniforme.

232. On procède de même que pour la moyenne arithmétique simple lorsque, au lieu de nombres absolus, on doit effectuer les calculs sur des nombres proportionnels exprimant combien de parties du tout représente chaque groupe de fréquences. On commence par déterminer l'importance de la classe, si elle est comprise entre deux limites, et on multiplie les nombres proportionnels par l'indice de la classe; le total des produits partiels divisé par 100 donne le chiffre du salaire moyen.

Exemple 11. — **Salaires moyens des ouvriers mâles adultes dans les mines de houille, en Belgique, au 31 octobre 1896**

Taux des salaires	Proportion °/₀ des ouvriers	Produit
Francs.		
1.00	0.25	0.250
1.745	1.40	2.443
2.245	4.67	10.484
2.745	12.50	34.312
3.245	26.85	87.128
3.745	21.94	82.165
4.245	18.33	77.810
4.745	8.25	39.146
5.245	3.08	16.155
5.745	1.28	7.354
6.245	0.71	4.434
6.745	0.31	2.091
7.245	0.43	3.115
	100.00	366.887

$$M = \frac{366 \text{ fr. } 88}{100} = 3 \text{ fr. } 67 \ (1)$$

(1) A titre de vérification, cfr. Bosco (A.), *Lezioni di statistica*, p. 508, qui, opérant par le procédé ordinaire, a trouvé le même résultat.

233. Si l'on veut se faire une idée précise de la moyenne arithmétique, on ne doit pas perdre de vue qu'elle consiste essentiellement en un partage égal et compensatoire. La moyenne arithmétique est le résultat d'une opération, par laquelle on divise un total en un nombre donné de parties égales; il en résulte que la moyenne multipliée par le nombre de termes, ou de parties, reconstitue le total. Cette propriété appartient à la moyenne arithmétique; elle est éminemment caractéristique. Partant de ces principes élémentaires, nous pouvons en déduire de brèves considérations sur la sphère d'application et sur les conditions intrinsèques du calcul de la moyenne. Sa compétence est certaine et absolue chaque fois qu'il s'agit de trouver une expression générale, pour caractériser une série statique, c'est-à-dire une série de nombres qui ne marquent, les uns par rapport aux autres, que de faibles oscillations. De pareilles séries ne sont pas rares en démographie; le rapport des naissances masculines aux naissances féminines en fournit un exemple remarquable. L'usage de la moyenne arithmétique est indiqué également dans de nombreux relevés des sciences météorologique et astronomique.

Dans une série parfaitement régulière, la moyenne atteint une haute valeur représentative; elle est à la fois le centre de gravité de la série, le point d'intersection de la surface de la courbe en deux parties égales, la valeur la plus probable autour de laquelle les autres viennent se grouper d'une manière symétrique, d'autant plus nombreuses qu'elles sont plus proches de la moyenne.

234. Mais les courbes qui suivent la loi de la répartition des erreurs accidentelles sont rares; un grand nombre de phénomènes sociaux dessinent une courbe asymétrique, soit qu'ils n'obéissent pas à la loi des erreurs, soit qu'il s'agisse de séries à caractère dynamique ou indéterminé. Dans ces cas, la moyenne a le caractère d'une expression synthétique d'un caractère abstrait; elle ne correspond d'habitude à

aucune des manifestations réelles du phénomène et a pour objet de donner une idée abrégée d'un ensemble complexe. Dans ces cas, la compétence de la moyenne arithmétique s'atténue ou disparaît et il appartient au chercheur, se guidant d'après les circonstances, de faire choix entre les moyennes arithmétique, géométrique, harmonique ou contre-harmonique, etc.

235. Puisque la moyenne, en général, est une expression synthétique des éléments complexes de la série, il est indispensable qu'elle exprime un ensemble de causes aussi étroitement unies que possible ; sans cela, elle perd sa signification et les comparaisons entre différentes moyennes sont vides de sens. De là résulte une règle essentielle, à savoir que les unités composant la série doivent présenter le caractère le plus homogène possible. Si les phénomènes sont influencés visiblement par des causes différentes, la moyenne qu'on en tirera sera dénuée de signification scientifique véritable ; elle n'exprimera qu'un complexus inintelligible et ne se prêtera pas à des comparaisons logiquement déduites (1).

Le caractère homogène d'une série doit s'apprécier sous les trois aspects suivants : 1° la nature des données ; 2° l'époque à laquelle elles se rapportent ; 3° le lieu où elles ont été réunies.

Une série peut être homogène dans son ensemble et pourtant certains des termes qui la composent peuvent obéir à des tendances particulières, en sorte qu'il y aurait avantage à ne pas les comprendre dans la moyenne générale. Si l'on considère d'une part tous les couples mariés, d'autre part le nombre d'enfants nés en légitime mariage, il est évident que l'on se trouve dans des conditions générales favorables pour calculer la moyenne des enfants par mariage. Mais, comme le fait observer Zizek (2), des doutes

(1) Zizek, *Statistical average*, p. 65.
(2) Id., *loc. cit.*, pp. 67-68.

se sont élevés sur le point de savoir s'il ne vaudrait pas mieux éliminer les couples sans enfants et calculer la moyenne seulement sur le nombre de personnes mariées ayant au moins un enfant. La raison alléguée serait que l'absence d'enfants trahirait une tare physiologique chez les mariés et que, pour réunir des données comparables, il vaudrait mieux considérer uniquement des couples dont le caractère normal serait dénoté par l'existence d'enfants. Nous n'avons pas à nous prononcer sur le bien-fondé de cette opinion physiologique, mais il est incontestable que la moyenne serait plus exacte si elle était prise uniquement sur des couples ayant eu un ou plusieurs enfants. Il peut être extrêmement important de connaître séparément le nombre d'unions stériles et surtout de s'assurer si leur nombre relatif a une tendance à augmenter ou s'il reste stationnaire. On pourrait arriver à une précision plus grande encore en établissant des catégories parmi les couples mariés, d'après l'âge des conjoints au moment de leur union, et en calculant pour chacune de ces catégories le nombre moyen d'enfants par mariage. Evidemment, ces données ne peuvent être réalisées sous leur forme la plus parfaite que si le statisticien dispose de données originales; celui qui travaille sur des données déjà publiées est forcé de les utiliser dans l'état où il les trouve.

236. Lorsqu'on étudie la condition d'homogénéité, on est surpris des conditions rigoureuses que le respect de cette condition impose. La question des salaires, si féconde en applications statistiques, en fournit une illustration nouvelle. Les statisticiens qui commencèrent l'étude des salaires sous le rapport statistique, se bornèrent souvent à des indications générales portant sur de vastes catégories de travailleurs; on parlait habituellement du « salaire moyen de l'ouvrier mineur », sans distinguer entre le fond et la surface, du « salaire moyen de l'ouvrier de fabrique », ou de celui de « l'ouvrier de métier ». Ces moyennes générales

présentaient le grave inconvénient d'éclairer très peu le public quant aux conditions réelles de la classe ouvrière, car ces indications vagues ne correspondaient à aucune situation existante. Pour obtenir une vue d'ensemble des salaires répondant à la condition d'homogénéité, il convient de partager les salaires en un assez grand nombre de classes. Ainsi, on mettra à part les enfants et les adultes; les premiers ne gagnent pas un salaire comparable à celui des seconds; leurs forces sont moindres, leur capacité professionnelle est à ses débuts, souvent leur salaire est considéré comme un salaire d'appoint, etc. D'autre part, nous savons que le sexe est un facteur dont l'importance est démontrée en ce qui concerne la fixation du taux des salaires; il convient donc d'envisager séparément le salaire des hommes et celui des femmes parmi les travailleurs adultes. Ensuite, il serait du plus haut intérêt d'observer les salaires par groupes d'âge : l'homme dans la pleine possession de son énergie produit plus que celui dont les forces déclinent; quel est l'âge de pleine production? quelle est la courbe du salaire d'après l'âge? Les statistiques des salaires n'ont pas encore porté la lumière sur ce point (1); il serait

(1) La Statistique belge des accidents du travail a rangé les victimes d'accidents du travail d'après leur âge et d'après le salaire moyen sur lequel la réparation a été calculée. Voici les données générales pour tous les groupes d'industrie (sans les entreprises agricoles proprement dites).

	Salaire annuel moyen
De moins de 15 ans	Fr. 468.67
15 à 20 ans	761.68
20 à 30 ans	1,178.13
30 à 40 ans	1,285.08
40 à 45 ans	1,287.10
45 à 50 ans	1,229.69
50 à 55 ans	1,173.01
55 à 60 ans	1,140.56
60 à 70 ans	1,052.19
70 et plus	923.49

Il résulte de ces chiffres que le salaire est le plus élevé à partir de 30 ans et qu'il reste à peu près stationnaire jusqu'à 45 ans; à partir de cet âge, il diminue pour se réduire, à l'âge extrême de 70 ans et plus, au niveau des années de la jeunesse. (Office du travail de Belgique. *Statistique des accidents du travail*, année 1906, t. I, p. 635.)

cependant fort intéressant de trouver la réponse à ces questions et de construire des moyennes vraiment homogènes, sur ces bases. Mais il ne suffit pas encore de ces distinctions; il en est d'autres tout aussi importantes : l'industrie et la profession. C'est seulement en opérant des groupements de plus en plus étroits qu'on parvient à calculer des moyennes homogènes.

L'époque à laquelle les données ont été recueillies est aussi un facteur à considérer, de même que la durée de l'observation. Des salaires recueillis en janvier ne sont pas comparables à d'autres, même se rapportant à des catégories identiques de travailleurs, observés en juillet. L'influence des saisons est extrêmement sensible parmi de vastes catégories de travailleurs; la morte-saison et la saison de pleine activité impriment à leurs gains des oscillations fortement marquées. C'est pour la même raison qu'on ne peut se contenter d'une période d'observation trop courte; l'idéal serait évidemment d'obtenir la succession des salaires pendant une année entière.

Enfin, il est bien évident que le milieu réagit sur les conditions des phénomènes. Pour obtenir des données absolument comparables ou homogènes, il faudrait ne grouper que des phénomènes soumis aux mêmes influences de milieu, ou, dans la comparaison, pouvoir éliminer l'influence due aux milieux divers.

III. — Moyenne géométrique.

237. La définition de la moyenne géométrique est la suivante : c'est la valeur qui correspond à la racine du produit des termes, portée à un exposant égal au nombre des termes.

La moyenne géométrique de 24 et 33 est donc :

$$G = \sqrt{24 \times 33} = 28.1424946$$

Lorsqu'il y a trois termes, au lieu de la racine carrée, on calcule la racine cubique du produit des termes. Ainsi, pour 24, 33 et 6, on a :

$$G = \sqrt[3]{24 \times 33 \times 6} = 16.8134152.$$

La valeur de l'exposant augmentant avec le nombre de termes, on est rapidement conduit, par la méthode arithmétique, à des calculs longs et compliqués. Supposons qu'il faille tenir compte des variations des prix d'un certain nombre de marchandises, admettons qu'elles soient au nombre de 45 comme dans l'*Index-Numbers* de Sauerbeck, on se trouverait devant une tâche presque insurmontable : c'est pour cette raison que l'on substitue au calcul arithmétique la recherche au moyen de logarithmes et l'on a alors cette définition très simple :

La moyenne géométrique d'une série de termes est le nombre naturel qui correspond au logarithme moyen des logarithmes de chacun des termes, ce qu'exprime la formule :

$$log.\ G. = \frac{1}{N} \Sigma\ (log.\ X) \qquad (14)$$

238. Il n'est pas inutile de faire une application étendue de cette formule ; on n'en trouve que très peu d'exemples pratiques dans les traités de statistique et cette pénurie embarrasse souvent l'étudiant qui se demande dans quelles conditions il doit appliquer la formule.

Parmi des amandes achetées dans un magasin de la ville, 54 ont été prises au hasard et mesurées avec une extrême précision, au dixième de millimètre. Il s'agit de déterminer, parmi ces 54 mesures, quelle est la moyenne géométrique. Le calcul arithmétique exigerait que l'on fît le produit de 54 termes l'un par l'autre et que l'on prît la racine 54e de

ce produit, ce qui est impossible. Le calcul par les logarithmes se dispose comme suit :

Exemple 12.

Longueur des amandes	Logarithmes des nombres ci-contre	Longueur des amandes	Logarithmes des nombres ci-contre	Longueur des amandes	Logarithmes des nombres ci-contre
(1/10e de m/m)					
258	2.41161971	310	2.49136169	334	2.52374647
262	2.41830129	310	2 49136169	335	2 52504481
265	2.42324587	312	2.49415459	339	2.53019070
267	2.42651126	313	2 49554434	344	2.53655844
269	2.42975228	314	2.49692965	348	2.54157924
276	2.44090908	318	2.50242712	349	2.54282543
291	2.46389299	321	2.50550503	352	2.54654266
294	2.46834733	324	2.51054501	356	2.55145000
295	2.46982202	325	2.51188336	359	2.55509445
295	2.46982202	325	2.51188336	359	2.55509445
299	2.47567119	325	2.51188336	362	2.55870857
299	2.47567119	326	2.51321760	362	2.55870857
299	2.47567119	329	2.51719590	363	2.55990663
302	2.48000694	329	2.51719590	364	2.56110138
304	2.48287358	329	2.51719590	365	2.56229286
307	2.48713838	330	2.51851394	366	2.56348109
308	2.48855072	331	2.51982790	369	2.56702637
310	2.49136169	333	2.52244423	372	2.57054294

Σ log. = 134.34713436.

M log. = 2.48790989.

G = 307.503.

239. L'emploi de la moyenne géométrique est peu répandu. Il faut attribuer ce fait à la complication des calculs ; les personnes auxquelles l'usage des tables de logarithmes n'est pas familier, se perdent facilement dans le dédale des colonnes de chiffres, surtout lorsqu'il s'agit de trouver le logarithme de nombres de 6 ou 7 chiffres. Cette difficulté matérielle, qui ne résulte que de l'absence d'exercice, ne doit pas éloigner de la moyenne géométrique, ni surtout la faire condamner quand son emploi est reconnu préférable à la moyenne arithmétique.

La moyenne géométrique est, dans des cas nombreux, recommandée par une série imposante d'autorités : Stanley Jevons, un des plus brillants esprits de son temps ; Galton qui, après Quetelet, a rénové la statistique ; Edgeworth dont les travaux font autorité ; Bowley et Yule dont nous avons eu maintes fois l'occasion de citer et de suivre les remarquables travaux, ont fait ressortir avec une grande force d'arguments, les avantages particuliers de cette moyenne.

Stanley Jevons l'a surtout recommandée à propos des statistiques des prix (1) ; selon lui, la moyenne arithmétique aurait pour effet, en général, d'exagérer les prix qui ont augmenté, aux dépens de ceux qui ont baissé. Si une marchandise A, dit Jevons, augmente de 100 p. c., son index-numbers est porté à 200 ; si, en même temps, une marchandise B diminue de 50 p. c., son index-numbers est ramené à 50 ; la moyenne arithmétique pour A + B est donc :

$$\frac{200 + 50}{2} = 125;$$

la moyenne géométrique serait :

$$\sqrt[2]{200 \times 50} = 100,$$

c'est-à-dire que le niveau général des prix resterait in-

(1) Stanley Jevons, *A serious fall on the value of gold ascertained*, London, 1863.

changé, tandis que la moyenne arithmétique le ferait apparaître 25 p. c. plus élevé.

Dans un travail subséquent (1), publié en 1865, Stanley Jevons n'apporte pas d'argument nouveau en faveur de la moyenne géométrique et se borne à reproduire l'exemple que nous venons de rappeler, mais dans son remarquable ouvrage *The principles of science*, publié en 1892, il reprend et amplifie ses premières conclusions. D'après l'éminent économiste anglais, « dans tous les calculs qui visent à exprimer le quantum des progrès de la société, la moyenne géométrique devrait être employée (2) » et il donne cet exemple : « une quantité quelconque (houille, fer, commerce) ayant, en 100 ans, augmenté de 100 p. c., on se tromperait fort en disant qu'en moyenne, à la fin de chaque décade, la quantité primitive a augmenté de 10 p. c.; de cette façon, on arriverait en réalité, à 159 p. c. La vraie moyenne serait la moyenne géométrique

$$\sqrt[10]{2} \text{ soit environ } 1.07,$$

ce qui donnerait une augmentation de 7 p. c. environ tous les dix ans ».

Galton a également défendu l'emploi de la moyenne géométrique dans les statistiques sociales et de la vie. D'après Galton (3) on ne pourrait admettre comme exacte en ce qui concerne les sciences de la vie l'hypothèse selon laquelle les erreurs en plus et les erreurs en moins par rapport à la mesure exacte sont également probables et que, par conséquent, la moyenne arithmétique de ces mesures est la plus proche de la vérité. La valeur la plus pro-

(1) Stanley Jevons, « On the variation of prices and the value of currency since 1782 ». (*Journal of the Statistical Society*, 1865, June, p. 295, note.)

(2) Stanley Jevons, *The Principles of Science*, London, 1892, p. 362.

(3) Sir Francis Galton, « The geometric mean, in vital and social statistics » (*Proceedings of the Royal Society*, n° 198, 1879. — Réimpression dans *Natural Inheritance*, London, 1889, annexe E, p. 238.)

bable, dans ces cas, serait donnée par la moyenne géométrique. Galton pense également que la moyenne géométrique est à préférer à la moyenne arithmétique lorsqu'il s'agit de phénomènes de la vie qui ont des relations avec la sociologie, par exemple l'augmentation de la population. Les phénomènes sociologiques, comme les phénomènes de la vie, dit Galton, doivent, en général, être traités à l'aide de la moyenne géométrique (1).

MM. Edgeworth et Bowley ont appuyé ces considérations et ont fait remarquer, en outre, que la moyenne géométrique possédait de plus l'avantage de rendre à peu près indifférent le choix de l'année de base dans les statistiques des prix.

Toutes ces considérations n'ont pu rendre fort populaire la moyenne géométrique. Non seulement, son emploi assez difficile rebute les chercheurs, mais son caractère mathématique abstrait ne contribue pas peu, comme M. Yule (2) le remarque avec raison, à lui conserver le caractère d'une méthode d'exception.

240. Lorsque les différences entre les nombres de la série ne sont pas accusées, le résultat du calcul ne change guère, que l'on emploie la moyenne arithmétique ou la moyenne géométrique.

---

(1) Une curieuse application de la moyenne géométrique a été proposée par I. H. VAN THUNEN, dans son ouvrage fameux : *Der Isolirte Staat*. D'après LUIGI COSSA (*Histoire des doctrines économiques*, trad. française, Paris, 1899, p. 410), van Thunen « crut avoir déterminé le juste salaire dans la formule $\sqrt{a \cdot p}$ c'est-à-dire la racine carrée du produit que l'on obtient en multipliant la somme exprimant la valeur des choses nécessaires à l'entretien de l'ouvrier par celle qui indique la valeur des produits obtenus par son travail ». Ainsi, si l'entretien de l'ouvrier est représenté par la somme de fr. 3.00 et la valeur ajoutée au produit par fr. 8.00, on a : $\sqrt[2]{3 \cdot 8} = 4$ fr. 8989, ce qui représenterait le juste salaire. « Mais, ajoute Luigi Cossa, ces prémisses étaient arbitraires et insuffisantes. »

(2) YULE (G. U.), *An Introduction to the theory of statistics*. London, 1911, p. 124.

Les nombres 20, 21, 22 ont pour moyenne arithmétique :

$$\frac{20 + 21 + 22}{3} = 21.$$

Leur moyenne géométrique est :

$$\sqrt[3]{20 \cdot 21 \cdot 22} = 20\,9841150$$

expression qu'on doit traduire en pratique par 21, valeur de la moyenne arithmétique.

241. Cette remarque limite l'emploi de la moyenne géométrique aux cas dans lesquels les termes de la série diffèrent d'une manière assez sensible.

Se trouvent spécialement dans ce cas, les séries qui réunissent des données relatives à l'accroissement de la fortune publique ou de l'activité économique, dans lesquelles interviennent plusieurs éléments doués d'une vitesse variable. Dans un travail antérieur, nous avons essayé de caractériser les progrès économiques de la Belgique au cours de la période 1880-1908 (1). Les moyennes, dans cette étude, sont des moyennes arithmétiques; reprenant une partie de nos résultats — ceux relatifs aux échanges, — de 1895 à 1908, nous avons calculé, à côté de la moyenne arithmétique, la moyenne géométrique. Le tableau ci-après fait apparaître les différences entre les deux résultats.

---

(1) JULIN (Arm.), « The economic progress of Belgium from 1880 to 1908 ». (*Journal of the Royal Statistical Society*, February, 1911, pp. 251-313.)

EXEMPLE 13. — **NOMBRES PROPORTIONNELS POUR CHAQUE ANNÉE (1884 = 100)**

INDICES	1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908
ÉCHANGES														
Importations (commerce spécial)	106.3	111.0	115.3	120.6	132.2	131.9	130.0	138.7	149.8	155.0	166.6	184.2	203.7	183 0
Exportations (commerce spécial)	93 5	97 8	106.8	111.6	120.7	121 5	113 5	117.9	125.6	124.0	132 9	157 0	162.2	144.7
Navigation maritime	168.4	183.7	195.7	202.1	211.9	208.7	228 0	249 3	267.9	274.4	285.2	317.8	329.6	329.2
Transport de houille et de coke par voies navigables	150.6	133.2	171.1	172.4	164.5	175 9	157.2	174.4	203.6	211.6	221.6	220.6	226 3	216.4
Nombre de voyageurs sur les chemins de fer de l'Etat et des compagnies	147.6	156.2	153 9	188.9	191.6	206 6	208.1	212.1	221.4	228.7	243.3	251 6	266.5	248.4
Transports de marchandises par les chemins de fer de l'Etat et des compagnies	114.3	120.9	110.9	117.2	140 2	148 0	145.8	152 9	158 9	162.9	172.6	186 7	188 5	163.7
Recettes des chemins de fer de l'Etat et des compagnies	105 6	110 7	126.2	130.1	136.1	142.9	141.1	146.0	152 0	157 5	167.9	175 0	178.1	169.7
Recettes des postes	130.4	133.0	137.7	143 6	154.4	160.0	163 0	166.2	171 3	177.1	188.2	196.0	197.4	—
Nombre de télégrammes d'affaires	140.4	142.3	147.0	173.7	189 2	174 4	185 3	179.5	177.0	184 3	207 4	195 6	207.7	195.3
Montant des effets escomptés par la Banque Nationale	126.0	119.7	123 9	126.7	134.3	143 7	138.4	138.8	145.3	142.2	150 6	164 1	169.8	164.9
Effets encaissés par la poste	133 4	139.1	143.7	150.6	157.8	170.7	173.9	174 4	174.8	178.8	184.2	195.3	204 2	202.8
Bons de poste et mandats-poste	126.1	129 5	135 4	134.5	134 2	144.7	151.5	156 1	159.5	162.8	170.1	179 7	183.9	184.3
Taux moyen de l'escompte	121.7	114 5	109 6	108.9	81 9	77 8	102.2	109 6	101.5	109.6	104.5	84.3	50.9	93.8
Nombre de faillites	128.0	124 7	127.3	—	121 6	138 5	131 5	131.0	126 3	128.5	135.2	142.8	141 2	144.6
Prix des principaux produits agricoles	77.1	76.0	82.6	89.6	85.3	94.0	95.5	94 8	90.6	88.7	96.2	97.1	100.6	100.8
Moyenne arithmétique	124,62	126 15	132 47	140 75	143 73	148.89	151.0	156 11	161.90	165.74	175.1	183 19	187 37	181.54
Moyenne géométrique (1)	122	124	130	137	139	144	146	152	152	159	168	175	174	173

(1) La moyenne géométrique a été calculée sur les nombres proportionnels arrondis et le calcul de cette moyenne n'a pas été effectué jusqu'aux décimal .

D'après l'exemple qui précède, les résultats calculés d'après la moyenne géométrique et d'après la moyenne arithmétique sont assez distants les uns des autres. La raison en est que certains éléments de la moyenne ont pris un développement considérable durant la période envisagée; il en va ainsi du commerce extérieur (importations et exportations), de la navigation maritime, des transports de houille et de coke par voies navigables, du nombre de voyageurs transportés par chemin de fer, du montant des effets encaissés et escomptés, etc. La moyenne arithmétique est très sensible à ces accroissements, tandis que la moyenne géométrique — c'est l'une de ses propriétés essentielles — ne les enregistre qu'en en modérant l'intensité. Ceci nous fournit une indication nouvelle pour l'emploi de la moyenne géométrique; elle est le plus souvent inférieure à la moyenne arithmétique et n'est jamais plus grande que celle-ci (1).

242. Lorsque la série ne possède pas un caractère dynamique, ou que tous ses éléments progressent d'une manière sensiblement égale, — comme c'est le cas le plus habituel dans les séries de prix relevés dans les *Index-numbers,* — les résultats obtenus à l'aide de la moyenne arithmétique ou de la moyenne géométrique restent sensiblement égaux. Comme l'a fait remarquer M. Edgeworth (2), ceci est un corollaire de la proposition plus générale d'après laquelle aucune espèce de moyenne résultant d'un groupe d'observations ne peut différer beaucoup d'aucune autre espèce de moyenne. Comme mesure pratique, on peut adopter la recommandation faite par Bowley de contrôler la moyenne géométrique par l'usage de la moyenne arithmétique :

---

(1) Cfr. une démonstration par le calcul différentiel de cette propriété par M. le professeur Warren Milton Persons. (Zizek, *Statistical averages,* pp. 195-196, note.)

(2) Edgeworth, « A defence of Index-Numbers ». (*The Economic Journal,* 1896, p. 136.)

quand les deux résultats diffèrent d'une manière sensible, il y a lieu d'adopter la moyenne géométrique qui est plus correcte à raison de la faculté qu'elle possède de diminuer l'effet des nombres les plus importants de la série.

IV. — Moyenne harmonique.

243. La moyenne harmonique est, des trois moyennes classiques, celle dont l'emploi est le moins fréquent. On la définit en disant qu'elle est *la réciproque de la moyenne arithmétique des réciproques des termes.* Elle a pour expression générale la formule :

$$H = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \dots \frac{1}{a_n}} \qquad (15)$$

et, en ne considérant que deux termes : a_1, a_2 :

$$\frac{2}{\frac{1}{a_1} + \frac{1}{a_2}}$$

et, en réduisant les termes à leur dénominateur commun :

$$\frac{2}{\frac{a_1 + a_2}{a_1 \cdot a_2}}$$

d'où l'on a, en multipliant le numérateur et le dénominateur de la fraction par a_1, a_2 :

$$\frac{2\ (a_1\ .\ a_2)}{a_1 + a_2} \qquad (16)$$

mais cette formule ne trouve son application que dans le cas particulier où la moyenne doit être prise entre deux termes (1).

---

(1) MESSEDAGLIA, « Calcul des valeurs moyennes ». (*Annales de démographie internationale*, 1880, p. 397.)

La valeur numérique de la moyenne harmonique est la plus basse; la plus élevée est celle de la moyenne arithmétique; la moyenne géométrique tient le milieu:

Rappelons l'application que nous avons faite (Cfr. *supra* n° 222) sur les nombres : 34, 38, 42, 46.

Moyenne arithmétique = 40.
Moyenne géométrique = 39,750
Moyenne harmonique = 39,498.

L'importance de l'écart entre les diverses moyennes dépend pour une large part, comme le fait remarquer M. Yule (1), de la grandeur de la dispersion par rapport à la grandeur de la moyenne.

244. Messedaglia qui, dans son mémoire « Calcul des valeurs moyennes », s'est occupé surtout de la moyenne harmonique, a fait remarquer le rapport qui existe entre la proportion et la moyenne arithmétique, d'une part, et la proportion et la moyenne harmonique d'autre part (2). La réciproque de chaque terme d'une proportion arithmétique ou par différence constitue une proportion harmonique des réciproques; d'où la réciproque de la moyenne arithmétique est la moyenne harmonique des réciproques, ou bien, la moyenne harmonique est la réciproque de la moyenne arithmétique des réciproques.

Le même auteur a fait remarquer que la moyenne harmonique est de nature à rendre des services dans l'appréciation de la valeur de la monnaie, c'est-à-dire de la quantité de marchandises qu'il est possible de se procurer avec une certaine somme d'argent (3). En général, on procède en prenant la moyenne arithmétique des prix; en cas de hausse de prix, la quantité de marchandises qu'on pourrait se procurer avec une même somme diminue d'autant. Mais,

(1) Cfr. aussi Yule, *An introduction to the theory of statistics*, p. 129.
(2) Messedaglia, « Calcul des valeurs moyennes ». (*Annales de démographie internationale*, 1880, p. 391.)
(3) Id., *loc. cit.*, pp. 414-418.

fait observer Messedaglia, à la moyenne arithmétique des prix ne correspond pas précisément, en sens inverse, la moyenne arithmétique des quantités. La moyenne arithmétique des prix conduit pour les quantités à une valeur inférieure à la valeur réelle; à la moyenne arithmétique des prix correspond la moyenne harmonique des quantités; si l'on veut se contenter d'une approximation déjà très proche, on peut prendre la moyenne géométrique.

V. — Principales propriétés mathématiques des moyennes.

245. Les moyennes possèdent des propriétés mathématiques importantes; elles sont à la base de plusieurs démonstrations qu'on trouvera plus loin. Nous passerons en revue les principales d'entre elles.

Moyenne arithmétique (M). — Avant de passer à l'examen des propriétés mathématiques de cette moyenne, il y a lieu de rappeler sommairement quelques caractéristiques résultant de sa nature : *a*) la moyenne arithmétique est le centre de gravité de la série, ce qui résulte de l'idée de partage égal qui se trouve à la base de sa définition; *b*) si les termes de la série sont égaux, la moyenne est égale à chacun d'eux; *c*) la moyenne, non plus que le résultat de l'addition, ne change pas, quel que soit l'ordre dans lequel les termes qui la composent se trouvent disposés.

246. *La somme des écarts positifs des termes par rapport à la moyenne est égale à la somme des écarts négatifs, c'est-à-dire que la somme algébrique des écarts est zéro.*

On appelle écart la différence numérique entre un terme de la série et la moyenne. Les écarts qui accusent une différence en plus sont les écarts positifs; ceux qui montrent une différence en moins, les écarts négatifs.

M. U. Yule (1) donne de cette propriété une démonstration très facile et très claire qui a l'avantage de justifier

(1) Yule, *loc. cit.*, pp. 110 et 114.

le procédé abrégé ou indirect du calcul de la moyenne que nous avons exposé plus haut. (Cfr. n° 229.)

Soient : X la valeur de chaque terme; A, une valeur arbitraire choisie pour représenter la moyenne; ξ, l'écart entre X et A; f, les fréquences.

Nous avons :

$$X = A + \xi \tag{22}$$

d'où

$$\Sigma\,(f.\ X) = \Sigma\,(f.\ A) + \Sigma\,(f.\ \xi)$$

or, A étant une valeur constante :

$$M = A + \frac{1}{N}\,\Sigma\,(f.\ \xi) \tag{23}$$

Si M et A sont identiques, on a :

$$\Sigma\,(f.\ \xi) = 0. \tag{24}$$

On peut aussi donner la démonstration sous cette forme (1) :

Soit la formule générale de M :

$$M = \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n} \tag{8}$$

Appelons x_1, x_2, x_3 les écarts de chaque terme a_1, a_2, a_3 par rapport à M; il vient que :

$$\begin{aligned} x_1 &= a_1 - M \\ x_2 &= a_2 - M \\ x_3 &= a_3 - M \\ &\;\;\vdots \\ x_n &= a_n - M \end{aligned}$$

(1) Benini : *Principii di statistica*, p. 94, note. Voyez aussi Virgilh. *Statistica*, p. 82. Hœpli, Milan, 1911.

En faisant la somme des écarts, nous avons :

$$x_1 + x_2 + x_3 + \ldots\ldots + x_n = a_1 + a_2 + a_3 + \ldots\ldots + a_n - n.\,\mathrm{M}$$

ou, en substituant à M sa valeur :

$$x_1 + x_2 + x_3 + \ldots + x_n = a_1 + a_2 + a_3 + \ldots + a_n -$$
$$- \, n \frac{a_1 + a_2 + a_3 + \ldots\ldots + a_n}{n} =$$
$$= a_1 + a_2 + a_3 + \ldots + a_n - (a_1 + a_2 + a_3 + \ldots + a_n)$$

D'où il suit évidemment que :

$$x_1 + x_2 + x_3 + \ldots\ldots + x_n = 0$$

Appliquons cette règle à notre exemple 3 (Cfr. n° 225) (nombre des mariages en Angleterre et dans le pays de Galles, de 1891 à 1900).

EXEMPLE 14. M = 239,410.5

Années	Mariages	M.	Ecarts	
1891	226,526 —	239,410.5 = x_1	— 12,884.5	
1892	227,135	239,410.5 = x_2	— 12,275.5	
1893	218,689 —	239,410.5 = x_3	— 20,721.5	— 70,049.5
1894	226,449 —	239,410.5 = x_4	— 12,961.5	
1895	228,204 —	239,410.5 = x_5	— 11,206.5	
1896	242,764 —	239,410 5 = x_6	+ 3,353.5	
1897	249,145 —	239,410.5 = x_7	+ 9,734.5	
1898	255,379 —	239,410.5 = x_8	+ 15,968.5	+ 70,049.5
1899	262,334 —	239,410 5 = x_9	+ 22,923.5	
1900	257,480 —	239,410.5 = x_{10}	+ 18,069 5	

247. La deuxième propriété de la moyenne arithmétique peut s'exprimer ainsi : *la somme des carrés des écarts à la moyenne est un minimum par rapport aux carrés des écarts qui seraient calculés sur un terme de la série différent de la moyenne, ou sur un autre nombre quelconque.*

Cette proposition peut être démontrée à l'aide de l'algèbre élémentaire (1). A la notation employée ci-dessus, ajoutons seulement les symboles y_1, y_2, y_3 pour désigner les

(1) BENINI : *Principii di statistica*, p. 95, note.

nouveaux écarts, et z la différence entre la moyenne arithmétique et la valeur adoptée pour en tenir lieu.

Supposons qu'au lieu de la moyenne arithmétique 239,410.5 (ex. 7), nous utilisons le chiffre 240,000 (A), dans ce cas, nous aurions :

EXEMPLE 15.

$$\begin{array}{lcr}
y_1 \ (1891) = 226{,}526 - 240{,}000 & = & -13{,}474 \\
x_1 = \ldots\ldots\ldots & = & -12{,}884.5 \\
\hline
z = \ldots\ldots\ldots & = & -589.5 \\
\mathrm{M} - \mathrm{A} = 239{,}410.5 - 240{,}000 & = & -589.5 \\
y_2 \ (1892) = 227{,}135 - 240{,}000 & = & -12{,}865 \\
x_2 = \ldots\ldots\ldots & = & -12{,}275.5 \\
\hline
z = \ldots\ldots\ldots & = & -589\ 5 \\
\mathrm{M} - \mathrm{A} = 239{,}410.5 - 240{,}000 & = & -589.5
\end{array}$$

etc., d'où il suit à l'évidence que z est une valeur constante par rapport à la différence entre la moyenne exacte et le terme qui en tient lieu et dont le signe est positif ou négatif selon les cas. On aura donc :

$$\begin{array}{l}
y_1 = x_1 + z \\
y_2 = x_2 + z \\
y_3 = x_3 + z \\
\ \cdot \quad \cdot \quad \cdot \\
\ \cdot \quad \cdot \quad \cdot \\
\ \cdot \quad \cdot \quad \cdot \\
y_n = x_n + z
\end{array}$$

Elevons au carré les égalités précédentes et faisons-en la somme :

$$\begin{array}{l}
y = x_1^2 + z_2 + 2\,x_1\,z \\
y_2^2 = x_2^2 + z_2 + 2\,x_2\,z \\
y_3^2 = x_3^2 + z_2 + 2\,x_3\,z \\
\ \cdot \quad \cdot \quad \cdot \quad \cdot \\
\ \cdot \quad \cdot \quad \cdot \quad \cdot \\
\ \cdot \quad \cdot \quad \cdot \quad \cdot \\
y_n^2 = x_n^2 + z_2 + 2\,x_n\,z \\
\hline
\Sigma\,(y^2) = \Sigma\,(x^2) + nz^2 + 2\,z\,\Sigma_{(}x)
\end{array}$$

Nous avons vu précédemment que $\Sigma\ (x) = 0$. Donc, le produit $2\ z\ \Sigma\ (x)$ est aussi égal à zéro. L'égalité précédente se réduit donc à cette forme :

$$\Sigma\ (y^2) = \Sigma\ (x^2) + n\ z^2$$

En élevant x au carré (x^2), nous avons rendu positifs les écarts qui, précédemment, étaient négatifs. L'expression $\Sigma\ (y^2) = \Sigma\ (x^2) + n\ z^2$ signifie donc que la somme des carrés des écarts de chaque terme calculés sur un nombre différent de la moyenne arithmétique est toujours plus grande que celle qui résulte de la comparaison avec la moyenne arithmétique : celle-ci réalise donc seule la condition du *minimum* (1).

248. On a aussi donné une autre démonstration simple de cette propriété. Nous la reproduisons à cause surtout de sa brièveté et de sa clarté. Appelons y les déviations mesurées d'après une autre valeur quelconque que la moyenne ; donnons la dénomination de d à la différence entre cette valeur et la moyenne, désignée par x. Alors, pour chaque terme de la série (x) (2), nous avons :

$$y = x - d;\ y^2 = x^2 + 2\ xd + d^2.$$

En portant les écarts au carré :

$$\Sigma\ (y^2) = \Sigma\ (x^2) - 2\ \Sigma\ xd + \Sigma\ d^2 = \Sigma\ (x^2) + nd^2.$$

Or $\Sigma\ xd = 0$ puisque l'on a démontré que les écarts positifs et négatifs s'annihilent ; donc $\Sigma\ y^2$ est le plus petit quand $nd^2 = 0$, c'est-à-dire quand $d = 0$.

249. L'exemple 14, d'après ce qui précède, donnerait les résultats suivants :

(1) La même démonstration se fait en utilisant le calcul différentiel. On en trouvera un exemple dans GABAGLIO, *Teoria della statistica*, t. II, pp. 119-120.

(2) HOOKER, « An Elementary explanation of correlation ». (*Quarterly Journal of the Meteorological Society of London*, 1908, p. 281.)

EXEMPLE 16.

Années	Mariages		Ecarts	X_2
1891	226,526	x_1	— 12,884.5	166,010,340.25
1892	227,135	x_2	— 12,275.5	150,687,900.25
1893	218,689	x_3	— 20,721.5	429,380,562.25
1894	226,449	x_4	— 12,961.5	168,000,482.25
1895	228,204	x_5	— 11,206.5	125,585,642.25
1896	242,764	x_6	+ 3,353.5	11,245,962.25
1897	249,145	x_7	+ 9,734.5	94,760,490.25
1898	255,379	x_8	+ 15,968.5	254,992,992.25
1899	262,334	x_9	+ 22,923.5	525,486,852.25
1900	257,480	x_{10}	+ 18,069.5	326,506,830.25
M =	239,410.5		+ 70,049.5	$\Sigma(x^2)$=2,252,658,054.50
			— 70,049.5	

La moyenne des carrés = 225,265,805.45.

Edgeworth a proposé de donner à cette moyenne de carrés le nom de « fluctuation ». Le symbole qui la désigne est (1) :

$$\mu^2 = \frac{\Sigma (x_1 - m)^2}{m} \qquad (25)$$

250. Une troisième propriété de la moyenne arithmétique peut être formulée en ces termes : *Si, pour un groupe de phénomènes, on possède plusieurs séries séparées d'observations, on peut, si le nombre des observations est identique, trouver la moyenne de l'ensemble du groupe à l'aide des moyennes des séries particulières; on peut également calculer la moyenne générale à l'aide des moyennes partielles quand les observations ne sont pas identiques, à la condition d'affecter chaque moyenne d'un poids égal au nombre des grandeurs qui entrent dans sa composition* (2).

En effet, on démontre que si N_1 est le nombre d'observations de la première série, N_2 celui de la seconde série, M_1

(1) MARCH (L.), « Essai sur un mode d'exposer les principaux éléments de la théorie statistique ». (*Journal de la Société de Statistique de Paris*, 1910, p. 450.)

(2) YULE (U. G.), *An introduction to the theory of statistics*, p. 115.

la moyenne dans le premier cas, M_2 la moyenne dans le second, on a (1) :

$$N.M = N_1 . M_1 + N_2 . M_2 \qquad (26)$$

Soit l'application suivante :

1re série		2e série	
236		244	
195	$M = \frac{862}{4} = 215,5$	199	$M = \frac{888}{4} = 222$
223		233	
208		212	
862		888	

Moyennes des deux séries

$$M = \frac{862 + 888}{8} = 218,75$$

ou

$$M = \frac{215,5 + 222}{2} = 218,75$$

251. La moyenne de carrés désignée plus haut (Cfr. n° 249) par le symbole μ^2 présente elle-même d'intéressantes particularités, nous retracerons deux d'entre elles d'une façon sommaire (2).

A. « *La moyenne des carrés de grandeurs associées est égale à la fluctuation de ces grandeurs augmentée du carré de leur moyenne.* » (March.)

Soient les nombres 32, 36, 40, 44 (a_1, a_2 a_3, a_4).

$$M = 38$$

$$\frac{\Sigma\,(a_1 + a_2 + a_3 + a_4)^2}{n} = 1464$$

$$M^2 = 1444$$

$$\mu^2 = 20$$

Donc : $\dfrac{\Sigma\,(a_1 + a_2 + a_3 + a_4)^2}{n} = M^2 + \mu^2 = 1464$

B. « *La fluctuation est égale au carré moyen des écarts, par rapport à une base quelconque, diminuée du carré de l'intervalle entre cette base et la moyenne.* » (March.)

(1) YULE, *op. cit.*, p. 115.

(2) Elles ont été démontrées avec une grande précision par L. MARCH, dans son « Essai sur un mode d'exposer, etc. ». (*Journ. Soc. Stat.*, Paris, 1910, p. 451 et suivantes.)

Soient les nombres comme ci-dessus et la valeur arbitraire 45 (A). Nous savons que M = 38 et $\mu^2 = 20$. Appelons x_1, x_2, x_3, x_4, les écarts des nombres par rapport à 45.

$$\frac{\Sigma (x)^2}{n} - (M - A)^2 = \mu^2$$

$$\frac{\Sigma (x)^2}{n} = 69$$

$$(M - A)^2 = 49$$

$\frac{\Sigma (x)^2}{n} - (M - A)^2 = 69 - 49 = 20$, valeur de la fluctuation.

252. *Moyenne géométrique.* — Cette moyenne présente deux propriétés particulièrement intéressantes.

En premier lieu, on remarque que *le carré de la moyenne géométrique coïncide avec la moyenne géométrique des carrés des termes.* Cette loi est susceptible d'une généralisation importante, car la relation exprimée plus haut est exacte aussi à l'égard des expressions d'une puissance plus élevée.

Cette propriété se démontre par la simple lecture des termes. En effet :

soient a_1, a_2 les termes de la moyenne, on a :

$$\left(\sqrt{a_1 \cdot a_2}\right)^2 = \sqrt{a_1^2 \cdot a_2^2}$$

et, à un degré plus élevé :

$$\left(\sqrt{a_1 \cdot a_2}\right)^3 = \sqrt{a_1^3 \cdot a_2^3}$$

ou enfin quel que soit le nombre des termes :

$$\left(\sqrt{a_1 \cdot a_2}\right)^n = \sqrt{a_1^n \cdot a_2^n}$$

Faisons application de cette propriété aux nombres 24 et 33 (Cfr. n° 237).

$$G = \sqrt[2]{24 \,.\, 33} = 28.1424946$$

Le carré de ce nombre (28.1424946) = 791.996.

Les carrés des termes sont : 576 et 1089.

La moyenne géométrique de ces nombres est : 792 alors que le carré de la moyenne géométrique est, comme nous venons de le voir, 791.996, c'est-à-dire 792 à quatre millièmes près.

253. La seconde propriété de la moyenne géométrique est la suivante : *La réciproque de la moyenne géométrique coïncide avec la moyenne géométrique des réciproques.*

La réciproque de la moyenne géométrique

$$\sqrt{a_1 . a_2}$$

s'écrit :

$$\frac{1}{\sqrt{a_1 . a_2}}$$

et la réciproque des termes :

$$\frac{1}{a_1} \text{ et } \frac{1}{a_2}$$

D'où l'on a la moyenne géométrique des réciproques :

$$\sqrt{\frac{1}{a_1} . \frac{1}{a_2}}$$

De là, se tire l'égalité :

$$\frac{1}{\sqrt{a_1 . a_2}} = \sqrt{\frac{1}{a_1} . \frac{1}{a_2}}$$

254. Enfin, *le logarithme de la moyenne géométrique est égal à la moyenne arithmétique des logarithmes des termes,* car nous avons évidemment :

$$G = \frac{\log. a_1 + \log. a_2 + \log. a_3 \ldots + \log. a_n}{n} \text{ ou}$$

$$\log. G = \frac{1}{N} \Sigma (\log. a_n) \qquad (14)$$

Les propriétés susdites de la moyenne géométrique ne sont pas applicables à la moyenne arithmétique, sauf le cas où il y a égalité entre les termes.

255. Nous avons déjà fait observer (Cfr. n° 243) que des trois moyennes classiques (arithmétique, géométrique, harmonique) la plus forte est la moyenne arithmétique, la plus faible la moyenne harmonique et que la moyenne géométrique se place entre les deux. Les statisticiens italiens, particulièrement Messedaglia (1) et Gabaglio (2) se sont attachés à faire ressortir les relations qui existent entre les différentes moyennes; nous empruntons à ce dernier auteur les considérations ci-après (3), en les faisant suivre d'un exemple.

Soient les trois nombres 4, 5, 6.

1° *La moyenne arithmétique est la moyenne arithmétique entre la moyenne harmonique* (valeur la plus basse) *et la moyenne contre-harmonique* (valeur la plus élevée).

Calculée sur les nombres ci-dessus :

$$M = 5$$

$$H = 4.865$$

$$C.\ H = 5.133$$

$$\frac{H + C.\ H}{2} = 4.999 \text{ ou } 5, \text{ valeur de } M.$$

2° On tire de là que *le total de la moyenne harmonique et de la moyenne contre-harmonique est égal à deux fois la moyenne arithmétique.*

$$H\ (4.865) + C.\ H.\ (5.133) = 9.998 = 2M \qquad (5)$$

Donc, étant donné deux quelconques des trois moyennes, on peut en déduire la troisième (4).

---

(1) MESSEDAGLIA, « Calcul des valeurs moyennes ». (*Annales de démographie internationale*, t. IV, 1880, p. 387 et suivantes.)

(2) GABAGLIO (A.), *Teoria generale della statistica*. Milano, 1888, t. II, pp. 206-211.

(3) Pour la démonstration mathématique, voir GABAGLIO, *loc. cit.*, pp. 206-211.

(4) MESSEDAGLIA, « Calcul des valeurs moyennes », *loc. cit.*, p. 394.

3° *La moyenne géométrique est la moyenne géométrique entre la moyenne harmonique et la moyenne arithmétique; chacune de ces deux dernières est une troisième proportionnelle de même nature, entre l'autre et la géométrique.*

$$G = \sqrt[3]{4 \,.\, 5 \,.\, 6} = 4.932$$
$$M = 5$$
$$H = 4.864887$$

Donc :

$$G = \sqrt{5 \,.\, 4,864887} = \sqrt[2]{24.324} = 4.932$$

4° Par conséquent, *en multipliant la moyenne arithmétique par la moyenne harmonique, on obtient le carré de la moyenne géométrique.*

$$M = 5$$
$$H = 4.864887$$
$$M \times H = G^2 = 4.932^2 = 24.324$$

5° *La moyenne arithmétique des réciproques est la réciproque de la moyenne harmonique* (1).

$$\frac{1}{H} = \frac{0,616666}{3} = 0,2055$$

$$H = 4.865, \text{ or } \frac{1}{4.865} = 0,2055$$

6° *La réciproque de la moyenne harmonique est la moyenne arithmétique des réciproques.*

Cette proposition découle de la précédente.

7° *La moyenne harmonique est la réciproque de la moyenne arithmétique des réciproques.*

La valeur des réciproques de 4, 5, 6 est :

$$4 = 0.250.000$$
$$5 = 0.200.000$$
$$6 = 0.166.666$$

---

(1) MESSADAGLIA, *op. cit. (loc. cit.)*, p. 391.

D'où la moyenne M = 0,205555.

$$H = 4.865, \text{ et } \frac{1}{4.865} = 0.2055$$

8° Enfin, il est clair que *si l'on divise la moyenne contre-harmonique par l'harmonique et si l'on multiplie le quotient par la moyenne harmonique, on obtient comme produit la moyenne contre-harmonique.*

$$\frac{C.H}{H} = \frac{5.133}{4.865} = 1,055$$

$$1,055 \times 4,865 = 5,132575$$

VI. — La médiane.

256. *Quand on range une série de variables d'après leur grandeur, la variable qui occupe la position centrale porte le nom de médiane.* C'est donc, comme dit M. Bowley, « la grandeur qui appartient à la variable qui se trouve à mi-chemin de la série (1) » ou, comme dit Fechner, « la grandeur au-dessus de laquelle et au-dessous de laquelle il y a un même nombre de termes, en sorte qu'elle divise la série par le milieu (2) ». Notre définition est à peu près d'accord avec celle de M. Yule (3) qui nous semble la plus claire. Avec le même auteur, nous ajouterons que, dans une construction graphique reproduisant la courbe des fréquences (voir ci-dessus ce qui est dit de la distribution des fréquences, chapitre premier, III), la médiane coupe le graphique par une verticale de façon à diviser l'aire de la courbe en deux parties égales.

Pour déterminer l'emplacement de la médiane, il faut procéder comme suit :

Si le nombre de variables comprises dans la série est un nombre impair, il suffit de laisser tomber à droite et à

(1) *Elements of statistics*, 2e édition, p. 124.
(2) *Kollektivmasslehre*, p. 13.
(3) *An introduction to the theory of statistics*, p. 116.

gauche un nombre égal de variables et d'attribuer la valeur médiane à la grandeur, représentée par une seule variable, placée entre les deux groupes ainsi constitués; un nombre impair de termes, ainsi divisé en deux parties, peut s'écrire $2n - 1$, et le $(n - 1)^e$ terme dans l'ordre des grandeurs est la seule variable dont la grandeur répondra à la définition de la médiane.

Si le nombre de variables est pair, la médiane tombe entre les deux groupements égaux qui partagent la série. Lorsque les deux grandeurs voisines sont les mêmes, ce qui peut se présenter fréquemment, car dans une série de variables classées d'après leur grandeur on observe souvent des répétitions, il n'y a point de difficulté; mais si les deux grandeurs voisines sont différentes, on en fait la moyenne et c'est à cette grandeur moyenne que se place la médiane.

Au lieu de calculer la moyenne des grandeurs-limites des deux parties qui composent la série entière, on peut aussi recourir à un procédé simple d'interpolation. Le recours à ce procédé de calcul est inutile lorsqu'il s'agit de variables rangées l'une après l'autre d'après l'ordre de leur grandeur, comme dans l'exemple donné au numéro suivant (longueur de 94 amandes classées, une à une, dans un ordre croissant). Mais les exemples d'un autre ordre sont plus fréquents en statistique. D'habitude, les statistiques rangent les fréquences dans des classes d'un certain intervalle, comme des classes de salaires, par exemple (salaires de 2 francs à fr. 2.49, de fr. 2.50 à fr. 2.99, etc.). Si le statisticien pouvait utiliser le matériel original, il lui suffirait de compter, à l'intérieur de la classe où se trouve la médiane, un nombre d'unités suffisant pour atteindre exactement la moitié du nombre total formant la série et le salaire médian. Malheureusement il n'en est pas ainsi dans le cas le plus ordinaire et il faut se résoudre à interpoler. L'interpolation à laquelle on a recours est basée sur l'hypothèse que la distribution des fréquences à l'intérieur de

la classe est uniforme, mais ceci n'est qu'une hypothèse qui n'échappe pas toujours à des objections fondées : la distribution des fréquences dans des classes assez étendues n'est certainement pas uniforme, ainsi qu'on peut le vérifier dans la statistique des salaires selon que les classes de salaires sont formées de divisions par vingt-cinq ou par cinquante centimes.

257. Pour exposer clairement le procédé d'interpolation auquel il vient d'être fait allusion, il y a lieu de recourir à un exemple.

Exemple 17.

Salaires de 71,955 ouvriers mâles adultes travaillant au fond, en mai 1900, dans 65 charbonnages situés en Belgique.

Taux des salaires	Nombre d'hommes adultes du fond ayant touché les salaires ci-contre	Taux des salaires	Nombre d'hommes adultes du fond ayant touché les salaires ci-contre
Moins de 1.50	8	de 9.50 à 9.99	240
de 1.50 à 1.99	119	10.00 à 10.49	182
2.00 à 2.49	642	10.50 à 10.99	75
2.50 à 2.99	1,492	11.00 à 11.49	55
3.00 à 3.49	3,084	11.50 à 11.99	16
3.50 à 3.99	5,706	12.00 à 12.49	24
4.00 à 4.49	12,077	12.50 à 12.99	14
4.50 à 4.99	11,850	13.00 à 13.49	10
5.00 à 5.49	7,716	13.50 à 13.99	7
5.50 à 5.99	6,495	14.00 à 14.49	1
6.00 à 6.49	6,061	14.50 à 14.99	1
6.50 à 6.99	5,865	15.00 à 15.49	2
7.00 à 7.49	5,047	15.50 à 15.99	0
7.50 à 7.99	2,612	16.00 à 16.49	1
8.00 à 8.49	1,318	16.50 à 16.99	1
8.50 à 8.99	771	17.00 à 17.49	1
9.00 à 9.49	461	20.00 à 20.49	1

Le nombre total des ouvriers dont les salaires ont été recueillis est de 71,955. Le nombre médian est 35,977. En additionnant les fréquences inscrites dans les huit premières classes, on arrive à un total de 34,978 ouvriers; la classe suivante est celle des salaires de 5 francs à fr. 5.49 et c'est dans cette classe que se trouve la médiane. Seulement, nous ignorons quel est le vrai chiffre du salaire médian; tout ce que nous pouvons dire, c'est qu'il est compris entre ces limites : 5 francs à fr. 5.49. Pour obtenir une approximation plus approchée, on procède de la sorte : la classe composée de 7,716 ouvriers étant trop vaste et dépassant la médiane il y a lieu de prendre de cette classe seulement le nombre d'ouvriers nécessaire pour atteindre le chiffre de 35,977, soit 999 ouvriers. D'autre part, il y a lieu de préciser l'intervalle de la classe et de prendre la valeur la plus élevée de la classe à laquelle on doit ajouter quelque chose.

On a donc :

$$4.99 + \frac{999}{7716} \cdot \frac{1}{2} \text{ ou } 0.0647 = 5.05$$

Le salaire médian calculé d'après l'interpolation ci-dessus serait donc de fr. 5.05; d'après la moyenne entre la valeur centrale des deux classes, le salaire serait de 5 fr. $\left(\frac{4.75 + 5.25}{2}\right)$. Le salaire moyen calculé en tenant compte du nombre des fréquences dans chaque classe, excepté dans la première où le salaire gagné par chacun des huit ouvriers qui y sont compris ne peut être déterminé, serait de fr. 5.357 ou fr. 5.36; la différence avec la médiane est somme toute sensible mais la médiane a, sur la moyenne, l'avantage d'être calculée très rapidement (1).

(1) Le calcul de moyenne ci-dessus exige 40 minutes, même avec l'aide de procédés mécaniques. Pour calculer la médiane, il faut à peine la dixième partie de ce temps.

258. La médiane est-elle une valeur purement abstraite, reposant sur une donnée conventionnelle, ou bien est-elle une mesure résultant de la nature des choses et conforme à une certaine loi de distribution par grandeur, quand on analyse un certain nombre de variables? La réponse à cette question est certes importante. On peut même y trouver un intérêt scientifique de premier ordre, car si les procédés statistiques peuvent montrer nettement la loi de distribution à laquelle obéissent les variables, ce résultat se rapproche sensiblement de cette conception fondamentale de la statistique que nous avons énoncée dans notre Introduction en disant que « le but de la recherche statistique consiste à exprimer ce que le phénomène a de permanent et de typique. On peut dire que la statistique, méthode propre aux phénomènes collectifs, a pour objet final la recherche de l'absolu parmi le relatif, du typique parmi l'accidentel, du permanent parmi le passager ».

Si l'on examine un à un, un grand nombre d'exemplaires d'objets d'une même espèce, comme les feuilles d'un arbre, ou des fruits d'une certaine sorte, on ne manquera pas de remarquer, comme le disent Palin et Ethel Elderton, que ces différents spécimens se différencient entre eux d'une manière sensible : parmi les feuilles arrachées à cet arbre, il en est de grandes et de larges, de petites et d'étroites; les fruits servis dans la corbeille qui orne la table ne sont pas d'égale grosseur. Mais il ne suffit pas de cette constatation banale: pour arriver à dégager une conclusion scientifique de ce fait, il faut procéder d'après une méthode spéciale dont nous ferons plus loin une application particulière.

Cette méthode consiste simplement, après avoir mesuré les variables selon tel ou tel de leurs caractères (longueur, épaisseur, largeur) à les ranger les unes à la suite des autres, dans un ordre croissant et à représenter au moyen de lignes verticales, placées à intervalles égaux, les différentes grandeurs obtenues pour chacune des variables. Sir Francis Galton a montré de la sorte les grandeurs d'une

série de cosses de pois en les rangeant dans un ordre croissant et en les photographiant dans cet ordre, toutes ensemble. Mais ce procédé ne peut s'appliquer utilement à toutes les variables, à cause de l'épaisseur des objets qui ne permettrait pas de discerner facilement l'allure de la courbe obtenue de la sorte, et d'habitude on a recours au procédé graphique que nous avons indiqué ci-dessus. Reprenant l'exemple donné par Palin et Ethel Elderton, nous avons nous-même procédé à une expérience qui nous semble intéressante et dont nous demandons au lecteur de suivre ici les résultats.

259. Nous avons acheté une demi-livre d'amandes (variété commerciale : « amandes princesses ») dans un magasin de Bruxelles; les amandes ont été enlevées d'un bocal contenant plusieurs kilogrammes, sans aucun choix; les amandes étaient au nombre de 94. Nous les avons mesurées avec soin au moyen d'un compas d'épaisseur, muni d'un vernier permettant de lire le dixième de millimètre. Toutes les mesures reproduites ci-après sont exprimées jusqu'au dixième de millimètre. Nous avons successivement mesuré la longueur des amandes, puis leur épaisseur.

En même temps que l'indication des longueurs, il était intéressant, après avoir déterminé la position de la médiane, de calculer les écarts de chaque variable à ce point.

C'est ce qui est réalisé dans le tableau suivant :

Exemple 18. — **Série des quatre-vingt-quatorze amandes.**

Numéros d'ordre	Écarts ξ	Longueurs	Numéros d'ordre	Écarts ξ	Longueurs	Numéros d'ordre	Écarts ξ	Longueurs
		Centimèt.			Centimèt.			Centimèt.
1	— 0.71	2.54	32	— 0.15	3.10	63	+ 0.09	3.34
2	— 0.67	2.58	33	— 0.15	3.10	64	+ 0.10	3.35
3	— 0.63	2.62	34	— 0.15	3.10	65	+ 0.14	3.39
4	— 0.60	2.65	35	— 0.13	3.12	66	+ 0.14	3.39
5	— 0.58	2.67	36	— 0.12	3.13	67	+ 0.18	3.43
6	— 0.58	2.67	37	— 0.11	3.14	68	+ 0.19	3.44
7	— 0.56	2.69	38	— 0.11	3.14	69	+ 0.23	3.48
8	— 0.54	2.71	39	— 0.10	3.15	70	+ 0.24	3.49
9	— 0.49	2.76	40	— 0.10	3.15	71	+ 0.27	3.52
10	— 0.44	2.81	41	— 0.07	3.18	72	+ 0.27	3.52
11	— 0.40	2.85	42	— 0.07	3.18	73	+ 0.29	3.54
12	— 0.34	2.91	43	— 0.04	3.21	74	+ 0.31	3.56
13	— 0.34	2.91	44	— 0.03	3.22	75	+ 0.34	3.59
14	— 0.33	2.92	45	— 0.03	3.23	76	+ 0.34	3.59
15	— 0.31	2.94	46	— 0.01	3.24	77	+ 0.37	3.62
16	— 0.30	2.95	Mé. } 47	0.00	3.25	78	+ 0.37	3.62
17	— 0.30	2.95	48	0.00	3.25	79	+ 0.38	3.63
18	— 0.30	2.95	49	+ 0.00	3.25	80	+ 0.39	3.64
19	— 0.26	2.99	50	+ 0.00	3.25	81	+ 0.39	3.64
20	— 0.26	2.99	51	+ 0.01	3.26	82	+ 0.40	3.65
21	— 0.26	2.99	52	+ 0.03	3.28	83	+ 0.41	3.66
22	— 0.26	2.99	53	+ 0.04	3.29	84	+ 0.42	3.67
23	— 0.25	3.00	54	+ 0.04	3.29	85	+ 0.44	3.69
24	— 0.23	3.02	55	+ 0.04	3.29	86	+ 0.47	3.72
25	— 0.21	3.04	56	+ 0.04	3.29	87	+ 0.49	3.74
26	— 0.20	3.05	57	+ 0.05	3.30	88	+ 0.49	3.74
27	— 0.20	3.05	58	+ 0.06	3.31	89	+ 0.50	3.75
28	— 0.18	3.07	59	+ 0.06	3.31	90	+ 0.57	3.82
29	— 0.18	3.07	60	+ 0.08	3.33	91	+ 0.57	3.82
30	— 0.17	3.08	61	+ 0.08	3.33	92	+ 0.71	3.96
31	— 0.15	3.10	62	+ 0.09	3.34	93	+ 0.71	3.96
						94	+ 0.75	4.00

$$\Sigma(\xi) = -1{,}260.$$
$$\Sigma(\xi) = +1{,}258.$$

Distribution de 94 amandes d'après leur longueur.

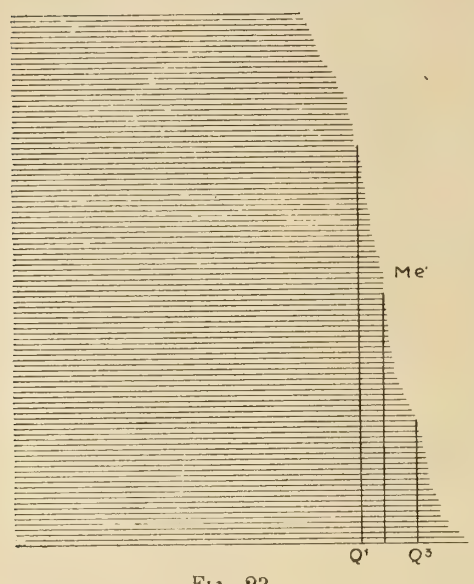


Fig. 23.

La régularité de la série est remarquable. La médiane se fixe aux n°s 47 et 48, endroit auquel correspond la longueur de 3 c. 25 mm.; la moyenne arithmétique est : 3.2498, soit un écart de deux dix-millièmes seulement avec la médiane; la somme des écarts positifs et négatifs par rapport à la médiane est respectivement de 1260 et 1258, exprimés en dixièmes de millimètre.

260. Maintenant, il serait intéressant de constater si cette médiane se trouverait fortement modifiée par l'introduction du hasard. Dans le but de faire cette constatation, nous avons tiré au sort 54 amandes, parmi les 94 qui ont servi à la première expérience : afin de conserver à l'élément hasard toute sa valeur, nous n'avons pas extrait les amandes elles-mêmes, mais nous avons tiré les numéros d'ordre correspondant à la première série au moyen de boules comportant une suite ininterrompue de 94 numéros.

Distribution de 54 amandes d'après leur longueur.

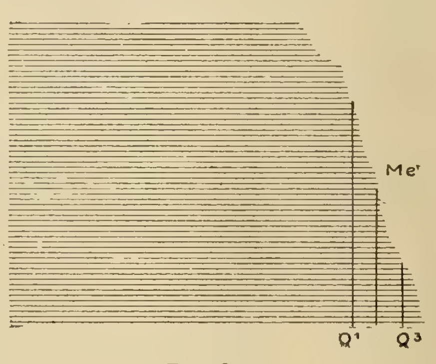


Fig. 24.

Voici les résultats de cette expérience.

EXEMPLE 19. — **Série des cinquante-quatre amandes tirées au hasard**

N° d'ordre (1)	Longueur	N° d'ordre (1)	Longueur	N° d'ordre (1)	Longueur
	Centimètres		Centimètres		Centimètres
2	2.58	33	3.10	62	3.34
3	2.62	34	3.10	64	3 35
4	2.65	35	3.12	65	3.39
6	2.67	36	3.13	68	3.44
7	2.69	38	3.14	69	3.48
9	2.76	41	3.18	70	3.49
12	2.91	43	3.21	72	3.52
15	2.94	46	3.24	74	3.56
16	2.95	Mé. { 47	3.25	75	3.59
18	2.95	48	3 25	76	3 59
19	2.99	50	3.25	77	3.62
21	2.99	51	3.26	78	3.62
22	2.99	53	3.29	79	3.63
24	3.02	55	3.29	81	3.64
25	3.04	56	3.29	82	3.65
28	3.07	57	3.30	83	3.66
30	3.08	59	3.31	85	3.69
31	3.10	61	3.33	86	3 72

La médiane a donc comme valeur 3 c. 25 mm., alors que la moyenne calculée sur les 54 termes se fixe à 3 c. 2225.

261. On peut faire un nouvel essai en tirant, au hasard, un nombre plus restreint de boules d'un jeu de loto, quarante-cinq par exemple, dont les chiffres correspondraient à ceux des numéros d'ordre des 94 termes composant la première série. Nous

Distribution de 45 amandes d'après leur longueur.

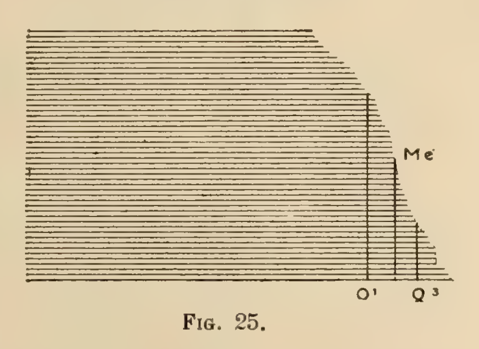


FIG. 25.

(1) Les numéros correspondent à l'ordre des grandeurs indiqué au tableau précédent (exemple 18).

avons fait cette troisième expérience, qui confirme entièrement les résultats des deux premières.

Voici les données :

Exemple 20. — **Série de quarante-cinq amandes tirées au hasard**

Nº d'ordre	Longueur	Nº d'ordre	Longueur	Nº d'ordre	Longueur
	Centimètres		Centimètres		Centimètres
1	2.54	40	3.15	66	3.39
2	2 58	42	3.18	67	3.43
4	2.65	44	3.22	68	3.44
7	2.69	46	3.24	69	3.48
10	2.81	47	3.25	70	3.49
11	2.85	49	3,25	72	3.52
13	2.91	50	3.25	75	3.59
15	2.94	Mé. 52	3.28	76	3.59
16	2.95	54	3.29	79	3.63
17	2.95	56	3,29	80	3.64
25	3.04	57	3.30	81	3.64
31	3.10	58	3.31	86	3.72
32	3.10	61	3.33	88	3.74
36	3.13	62	3 34	89	3.75
38	3.14	65	3.39	91	3.82

Nous obtenons comme valeur de la médiane 3 c. 28 mm.; la moyenne arithmétique donne comme valeur typique de la série 3 c. 2515.

Les données qui précèdent mettent bien en vue la valeur représentative de la médiane; les mesures simplement obtenues à l'aide de la médiane ne s'écartent pas sensiblement de celles résultant du calcul de la moyenne; de plus, elles ne sont pas fort différentes quoiqu'on opère sur la série entière ou sur des séries plus limitées dont les termes sont désignés au hasard par un tirage. On peut répéter des

expériences du même genre en prenant pour sujets les feuilles d'un arbre, le nombre de fèves ou de pois contenus dans une série de gousses, etc., etc. Les graphiques en regard permettent de suivre entièrement les résultats de l'expérience. Le lecteur ne manquera pas de remarquer combien l'allure de la courbe dessinée par la succession des variables reste identique dans les différents cas.

De même qu'on détermine la médiane par la position de la valeur centrale, on peut calculer la valeur des quartiles et des déciles placés à l'endroit correspondant au quart et au dixième du nombre des termes. L'usage des quartiles et des déciles sera exposé plus loin à l'endroit où il sera traité de la déviation type ou *standard deviation.*

262. Le calcul de la médiane exige un nombre raisonnable de termes, rigoureusement homogènes quant à leur objet. Si la série renferme un nombre de termes insuffisant, la médiane ne serait pas calculée dans de bonnes conditions. La répétition trop fréquente de la même mesure dans une série de variables est aussi une condition défectueuse pour la recherche de la médiane, bien que quelques répétitions ne puissent toujours être évitées.

Lorsqu'une série d'observations est formée de la réunion de deux séries partielles, la moyenne de la série complète peut être trouvée en combinant ensemble les moyennes séparées des deux séries particulières, c'est-à-dire en multipliant chaque moyenne séparée par le nombre de variables comprises dans la série, en additionnant les résultats partiels et en divisant le total par le nombre total de tous les termes. (Cfr. n° 250.)

Supposons que la série des 94 amandes soit formée de deux séries particulières constituées parmi la série générale en sélectionnant un groupe de 45 amandes et un autre

groupe de 49 amandes (1). Nous avons, d'après les résultats des variables comprises dans chaque série :

1° 45 amandes avec une moyenne de 3 c. 3335 (longueur moyenne) ;

2° 49 amandes avec une moyenne de 3 c. 1728 (longueur moyenne) ;

La moyenne générale est égale à :

$$94 . M = (45 \times 3.3335) + (49 \times 3.1728) = \frac{150.01 + 155.47}{94} = 3.249$$

moyenne qui résulte de la série complète, comme nous l'avons montré au n° 254 qui précède.

Mais nous ne pouvons procéder à un calcul semblable en ce qui concerne les médianes de nos deux séries particulières, car en multipliant la médiane par le nombre de termes de la série, on ne reconstitue pas, comme c'est le cas avec la moyenne arithmétique, le nombre total des unités comprises dans la série; toutefois, un calcul de l'espèce se rapprocherait d'autant plus de la moyenne que la courbe des deux séries particulières serait plus étroitement symétrique : nous avons déjà dit que dans une distribution

(1) Cette sélection a été faite tout naturellement et en suivant seulement les lois du hasard, car en achetant les amandes qui ont servi à notre expérience, nous avions eu soin de faire peser deux fois un quart de livre; la première série contenait 45 amandes, la seconde 49. Voici la répartition des longueurs pour chacune des séries :

PREMIÈRE SÉRIE

2,67	3,14	3,33	3,64
2,76	3,15	3,34	3.64
2,91	3,18	3,34	3,67
2,94	3,21	3,39	3,69
2,99	3,22	3,44	3,74
2,99	3,22	3,48	3,74
3,02	3,24	3,49	3,75
3,05	3,25	3,52	3,96
3,07	3,26	3,52	4,00
3,08	3,29	3,59	
3,12	3,29	3,62	
3,13	3,31	3,63	

DEUXIÈME SÉRIE

2,54	2,95	3,14	3,35	3,96
2,58	2,95	3,15	3,39	
2,62	2,99	3,18	3,43	
2,65	2,99	3,25	3,54	
2,67	3,00	3,25	3,56	
2,69	3,04	3,25	3,59	
2,71	3,05	3,28	3,62	
2,81	3,07	3,29	3,65	
2,85	3,10	3,29	3,66	
2,91	3,10	3,30	3,72	
2,92	3,10	3,31	3,82	
2,95	3,10	3,33	3,82	

idéalement symétrique, la moyenne, la médiane et le mode se fixent au même point; l'écart même de ces différentes mesures peut être pris comme un indice de la symétrie de la courbe elle-même. Ainsi, les deux séries particulières étant dans notre exemple à peu près symétriques, la mesure des deux médianes obtenues (3,25 et 3,28) donne une mesure type très rapprochée (3,265) de la médiane générale (3,250), mais qui ne se confond pas avec celle-ci.

263. La médiane ne peut être déterminée dans toute série quelconque, mais seulement dans celles qui groupent des données individuelles entre un certain nombre de classes, par exemple les salaires gagnés par les ouvriers d'une certaine profession, la stature des hommes d'un régiment, etc. La condition essentielle pour que la médiane puisse être calculée, est que les unités de la série soient rangées les unes à la suite des autres d'après leurs dimensions.

Lorsque la série est composée de nombres proportionnels ou de moyennes, l'usage de la médiane est interdit, parce que les poids correspondant à chaque donnée ne sont pas égaux entre eux. Les coefficients de mortalité calculés pour une série de groupes professionnels, peuvent, il est vrai, être rangés selon leur ordre d'importance, mais nous ne pouvons, sous cette forme, en déterminer le centre. « Le centre actuel de la série, dit Zizek, est situé plus haut ou plus bas que la donnée centrale, selon que les données inférieures ou supérieures concernent des quantités plus considérables et sont par conséquent de poids plus grand (1). » Cette observation ne vise, bien entendu, que des nombres proportionnels pris sur des données différentes ou qui se rapportent à plusieurs époques; elle ne s'applique pas aux nombres proportionnels calculés sur un ensemble de données homogènes : dans ce cas, les chiffres propor-

(1) Zizek, *Statistical averages*, p. 204.

tionnels conservent la même portée que les nombres absolus et peuvent être utilisés au même titre que ceux-ci.

En effet, en utilisant les pourcentages, au lieu des nombres absolus, dans l'exemple que nous avons reproduit plus haut (*salaires de 71,955 ouvriers mineurs Cfr. n° 257*) on obtient, à la classe de salaires fr. 4.50 à 4.99, une proportion totale de 48.61 p. c. d'ouvriers; il faut donc ajouter à cette proportion, 1.38 p. c. d'ouvriers appartenant à la classe suivante pour arriver aux 50 p. c. où se place la médiane; or :

$$4.99 + 1.38\ \% = 4.99 + 0.06 = 5 \text{ fr. } 05 \text{ c.},$$

résultat auquel on était déjà arrivé en utilisant les nombres absolus. (Cfr. n° 257.)

264. On peut résumer comme suit les avantages et les inconvénients des médianes. Au nombre des avantages de la médiane, on peut signaler :

1° La manière exacte dont on peut la localiser;

2° La facilité et la rapidité des calculs à l'aide desquels on la détermine, avantages qu'elle partage avec la moyenne arithmétique, en ce qui concerne la facilité et où elle se montre supérieure à la moyenne si l'on envisage la rapidité;

3° Mais ce qui rend la médiane précieuse, c'est surtout ce fait qu'il n'est pas nécessaire de connaître la valeur des termes extrêmes si l'on connaît leur nombre et si l'on sait que leur valeur ne dépasse pas une certaine limite. Or, ces conditions se présentent très fréquemment en statistique. Dans les statistiques de salaires, on ne donne généralement pas de détails sur les salaires inférieurs à une certaine somme, on se borne à indiquer le nombre d'ouvriers gagnant un salaire inférieur à tel chiffre (1); de même, on

(1) Les statisticiens devraient prendre pour règle de toujours donner les extrêmes et d'adopter un mode d'énumération complète. L'inconvénient qu'on fait valoir pour s'écarter de cette règle est le trop grand nombre de colonnes.

n'énumère pas, en général, tous les salaires, jusqu'au plus élevé, mais on donne le nombre global d'ouvriers recevant un salaire supérieur à tel chiffre. Or la médiane s'accommode de ces renseignements sommaires, tandis que la moyenne ne s'y prête pas. Le calcul de la moyenne sur ces bases rencontre des difficultés parce que l'une des classes, la dernière, est illimitée, de sorte qu'on ne peut simplement avoir recours à l'hypothèse d'après laquelle les salaires se répartiraient, par groupes égaux, entre les limites de la classe (1).

Il n'y a pas que les statistiques publiées dans ces conditions qui se prêtent à l'emploi de la médiane tandis qu'elles se refusent à celui de la moyenne; il y a encore à signaler la facilité que la médiane apporte à l'exécution de certaines enquêtes : dans une statistique des revenus, par exemple, ce sont les classes extrêmes qui sont surtout difficiles à déterminer : les très petits revenus, à cause de leur nombre, de leur variété et de leur incertitude; les très gros, à cause de la résistance que l'enquête rencontrera. Le revenu moyen ne pourrait être fixé avant d'avoir surmonté toutes ces difficultés; pour le revenu médian, il suffirait de connaître le nombre de revenus au-dessous et au-dessus d'un certain chiffre, ce qui peut être obtenu sans difficulté (nombre de revenus inférieurs à 1,000 francs, nombre de revenus supérieurs à 500,000 francs);

4° La moyenne se trouverait fortement influencée par la présence ou l'adjonction d'un terme très différent de ceux de la série; la médiane ne marquerait, dans ce cas, qu'une sensibilité fort relative. Supposons une série de 49 revenus

L'Office du Travail de Belgique a trouvé une solution à cette difficulté en publiant en note les salaires extrêmes qui se trouvaient totalisés dans la première et la dernière colonne.

(1) Pourtant, en admettant que la série suive la courbe de la loi asymétrique de Gauss, on peut déterminer la moyenne d'une série dont les extrémités sont illimitées. Mais, par la simplicité et la rapidité de la détermination du point caractéristique, l'emploi de la médiane est à préférer à tout autre.

montant ensemble à 490,000 francs; on a la moyenne $\frac{490,000}{49} = 10,000$ francs; à ces revenus s'ajoute celui d'un millionnaire possédant 500,000 francs de revenu; on a la nouvelle moyenne $\frac{490,000 + 500,000}{50} = 19,800$ francs. Au contraire, la médiane ne serait modifiée que de peu; avec 49 revenus, elle se fixerait à la classe correspondant au 25[e] revenu; l'adjonction du revenu du millionnaire obligerait seulement le statisticien à adopter, pour la médiane, la moyenne entre le 25[e] et le 26[e] revenu, soit une différence à peu près insignifiante avec le chiffre obtenu d'abord. Pour l'appréciation du revenu d'une collectivité, la médiane serait donc beaucoup plus exacte que la moyenne, car il importe peu à 49 possesseurs d'un revenu médiocre qu'à leur nombre vienne s'ajouter un seul millionnaire.

265. Parmi les inconvénients de la médiane, on peut signaler :

1° En multipliant la médiane par le nombre de termes, on ne reconstitue pas le total de la série et, par conséquent, l'addition de deux médianes relatives à deux séries particulières ne donne pas le même résultat que la recherche de la médiane opérée sur une série unique formée des mêmes unités, — théorème que nous avons démontré plus haut. (Cfr. n° 250.);

2° La médiane ne tient presque pas compte des variations extrêmes, de sorte que son emploi est limité aux cas où ces variations peuvent être négligées sans inconvénient. On préfèrera la médiane pour l'estimation du revenu d'une collectivité : on aura plutôt recours à la moyenne s'il s'agit d'établir un prix de revient;

3° Elle n'est pas applicable à toutes les séries, notamment à celles qui se composent de nombres proportionnels calculés sur des masses différentes, ou aux séries chronologiques, ou aux séries composées de moyennes;

4° On ne peut non plus rechercher la médiane qu'à la condition que les termes de la série puissent être rangés dans l'ordre de leur grandeur. Au contraire, l'ordre dans lequel sont placés les termes servant au calcul de la moyenne et les changements qui peuvent s'opérer dans cet ordre, n'ont aucune influence sur la moyenne.

Cependant, dans l'ensemble, la somme des avantages l'emporte sur celle des inconvénients et ceci est particulièrement exact en ce qui concerne l'étude de la répartition des revenus et des salaires.

VII. — La dominante (mode).

266. Le professeur Pearson a proposé, en 1895, de donner le nom de « mode » à *la valeur d'une série désignée par l'abscisse correspondant à l'ordonnée de fréquence maximum* (1). Cette valeur a d'abord été désignée par des auteurs français sous le nom de « valeur normale » ou simplement « normale », puis plus tard, par l'expression « la dominante »; les Italiens lui donnent le nom de « norma », les Allemands celui de « dichtestes Wert » (2). Il peut paraître avantageux de conserver le nom que lui a donné le professeur Pearson, afin de maintenir l'unité de terminologie, et pour le motif que le terme « mode » est employé par les statisticiens-mathématiciens de l'école anglaise dont les travaux doivent être constamment consultés et suivis dans la matière qui nous occupe. Mais, en français, l'expression « mode » n'a aucun sens qui se rapproche de celui que nous envisageons ici, et, sur la suggestion de M. Lucien March, nous avons décidé d'adopter l'expression « la dominante ».

Avant tout, il s'agit de définir la dominante d'une façon

---

(1) Pearson (K.), Skew Variation, etc., *loc. cit.*, p. 345, note.
(2) Zizek (F.), *Statistical averages*, p. 322.

un peu plus détaillée, de reconnaître exactement sa valeur, de se faire une idée des avantages qu'elle réunit.

La dominante est, de toutes les valeurs composant une série, celle qui se présente avec la fréquence la plus élevée et autour de laquelle les autres valeurs se groupent avec la densité la plus forte. Autrement dit, la dominante correspond à la classe la plus élevée de la série et les classes qui l'avoisinent à gauche et à droite ont une fréquence moins élevée qu'elle. Il suit de là que, dans une série régulière, reproduisant d'une manière plus ou moins parfaite la courbe normale des erreurs, la dominante se trouve au centre et le point auquel elle se fixe correspond à celui où se placent la moyenne et la médiane. Dans une courbe idéale, les trois points se confondraient en un seul, mais comme nous l'avons dit plus haut, il n'y a presque aucun phénomène physique ou économique dont la courbe reproduise exactement celle des erreurs et même ceux qui s'en rapprochent d'une façon sensible sont relativement rares ; il en est, au contraire, un grand nombre qui semblent obéir à une loi de distribution asymétrique rappelant, de plus ou moins loin, la courbe normale. Aussi n'est-il pas rare de constater dans une série la présence d'un grand nombre de sommets, alors que la courbe normale ou modérément asymétrique n'en contiendrait qu'un seul. Il a donc été nécessaire de trouver des méthodes qui permissent d'absorber les différents sommets en une série plus régulière, qui n'en contient qu'un seul; ces procédés seront exposés plus loin.

267. Tandis que la moyenne est une abstraction mathématique, la dominante correspond à l'idée générale que se font la majorité des gens lorsqu'on demande de définir une situation usuelle ou normale. Si l'on interroge un chef d'entreprise ou un dirigeant de syndicat au sujet du salaire moyen que gagne dans ses ateliers ou dans la profession représentée par le syndicat, un ouvrier mâle adulte,

ni l'un ni l'autre ne songera à procéder à un calcul lent et minutieux; si nous analysons leur travail mental, nous verrons que pour trouver la réponse à notre question, ils commenceront par ne pas tenir compte des jeunes ouvriers qui n'ont pas encore atteint le complet développement de leurs capacités physiques et techniques, et ils élimineront de même les vieux ouvriers qui ne jouissent plus de leur puissance de travail intégrale; ils ne songeront pas davantage aux salaires exceptionnellement élevés que gagnent quelques spécialistes ou des ouvriers d'une habileté exceptionnelle. Ils auront en vue le plus grand nombre des ouvriers qui répondent au type professionnel dans l'industrie en question et ils répondront : « en moyenne, l'ouvrier de telle industrie gagne x francs ». Le chiffre cité ne coïncidera probablement pas avec la moyenne, parce que celle-ci tient compte de tous les salaires, des plus faibles comme des plus élevés, mais le chiffre donné a des chances d'être assez près de la dominante, c'est-à-dire de la valeur de plus grande densité.

Il est permis de dire que la dominante se rapproche plus de la conception vulgaire de la « moyenne » que la moyenne arithmétique elle-même. Si nous considérons avec attention les procédés arithmétiques étudiés au cours de ce chapitre, nous voyons que leur utilité principale est de nous fournir la mesure typique d'un ensemble complexe de faits que nous voudrions évaluer au moyen d'une expression simple. Or, dans un grand nombre de cas, nous aurions avantage à connaître ce qui se présente le plus souvent, la mesure de plus grande densité, plutôt que le résultat d'un calcul qui n'est qu'une simple conception arithmétique. Dans un magasin de confections, le marchand se basera plutôt sur les mesures les plus fréquentes parmi les personnes qui composent sa clientèle que sur la moyenne proprement dite; le *quod plerumque fit* est pour lui une règle plus sûre que la moyenne, et ses prévisions se rapporteront à « la dominante » de préférence à toute autre conception synthétique.

268. Quant aux avantages que présente la dominante, on peut, avec MM. King et Bowley (1) les résumer comme suit :

1° La dominante est un procédé utile à employer chaque fois qu'il est désirable de ne pas tenir compte des variations exceptionnelles. — Parmi les ouvriers d'un atelier, en général, d'une capacité inférieure et ne gagnant qu'un salaire médiocre, il s'en rencontre par hasard quelques-uns d'une grande activité qui gagnent, en travaillant aux pièces, beaucoup plus que leurs camarades. Si l'on calcule le salaire moyen des ouvriers de cet atelier, on devra tenir compte des salaires exceptionnels de ces ouvriers d'élite et la moyenne en sera augmentée d'autant. Au contraire dans le calcul de la dominante, on cherchera à déterminer dans quel groupe de salaires on trouve le plus grand nombre de travailleurs, résultat sur lequel la présence de quelques ouvriers plus favorisés ne peut exercer d'influence.

Si l'on voulait comparer les salaires dans une même industrie, à deux époques différentes, peut-être serait-il préférable de choisir la dominante plutôt que la moyenne, car ce dernier résultat serait influencé par la présence de taux extrêmes, très bas ou fort élevés, qui peuvent exister ou disparaître sans modifier d'une manière sensible la condition économique de la masse des ouvriers. Le choix de la dominante répond donc à une objection faite fréquemment à l'emploi de la moyenne en ce qui concerne les salaires (2).

---

(1) King (W.), *Elements of statistical method*, New York, 1912, p. 125. — Bowley (A. L.), *Elements of statistics*, p. 123.

(2) M. Bowley propose (*loc. cit.*, p. 123) un exemple emprunté aux collectes qui se font dans les églises pour les besoins du culte et que nous nous permettons de développer quelque peu. Un curé de village, désireux de se rendre compte du produit des collectes faites aux différents offices du dimanche, ferait mieux de se baser sur la dominante que sur la moyenne; aux messes matinales, il constatera par exemple que l'offrande des fidèles est, le plus souvent, de la plus petite pièce de monnaie, tandis qu'aux messes qui se célèbrent plus tard et sont surtout fréquentées par des personnes aisées de la paroisse, les décimes sont en majorité. Supposons que parmi les hôtes occasionnels de ce village, se trouve un châtelain qui verse cinq ou

2° Pour déterminer la dominante d'une manière *approchée,* il n'est pas nécessaire d'avoir des données précises concernant les termes extrêmes de la série; il suffit de savoir qu'ils sont peu nombreux; — bien entendu, il faut qu'il s'agisse des termes extrêmes de la série —; on peut se contenter d'une approximation, ou d'un degré de précision qui permette de déterminer le maximum atteint par la fréquence. Par contre la dominante ne peut être connue d'une façon *exacte* que si tous les termes sont connus.

Dans une série de salaires, si nous ne voulons pas atteindre un haut degré d'exactitude, il ne faut pas connaître nécessairement le nombre d'ouvriers qui gagnent les salaires, rangés par classe, en dessous d'un certain taux, quand ce nombre est peu important, ni le nombre de ceux, dénombrés par classes, qui gagnent un salaire supérieur à un certain taux-limite. Puisque nous admettons *a priori* que leur nombre est peu important, il est impossible que la dominante change de place. Au contraire, pour déterminer le chiffre de la moyenne, la dernière précision est nécessaire. Cette facilité qu'offre le mode est précieuse, parce que les séries statistiques reproduites dans les publications officielles ne donnent que rarement les données précises jusqu'à la dernière classe mais elle est moins étendue qu'en ce qui concerne la médiane;

3° Enfin, la dominante donne une valeur réelle, la valeur la plus fréquente de la série, de façon qu'elle répond bien à ce que veut exprimer la notion vulgaire de moyenne. — Au

---

dix francs à la collecte de la grand'messe : la dominante ne sera pas changée, mais bien la moyenne; et si, par hasard, un jour de chasse, le châtelain assiste à la première messe et donne son offrande habituelle, la moyenne de la collecte sera changée encore plus, tandis que la dominante ne sera pas modifiée. Or, c'est sur une situation habituelle que le curé doit tabler pour le budget de son église et il fera bien de ne pas utiliser dans ce but la moyenne influencée par la générosité d'un hôte de passage; la dominante, c'est-à-dire le nombre habituel de paroissiens qui versent à la collecte 2, 5, 10 ou 50 centimes, lui fournira une base d'appréciation sensiblement plus juste.

contraire, la moyenne mathématique répond à une conception idéale, qui ne trouve peut-être aucune réalisation en pratique. Il se peut très bien que la moyenne du salaire d'un ouvrier, ou des appointements d'un employé ne corresponde au salaire effectif d'aucun ouvrier ou d'aucun employé compris dans la série, ce qui est assez déroutant pour qui ne se rend pas exactement compte de la notion de « moyenne ».

269. Il existe plusieurs procédés pour déterminer la position de la dominante.

Le calcul peut atteindre la grande précision si, connaissant tous les termes, on utilise les courbes de Pearson.

Mais, en général, on s'en tient à des procédés approximatifs. Nous résumerons en premier lieu les deux méthodes de M. Bowley, puis la formule de M. Pearson; quant aux procédés graphiques, nous en dirons un mot au chapitre consacré à la statistique graphique à l'endroit où nous parlerons des moyens de trouver la médiane.

Pour trouver la dominante (mode), on peut, d'après M. Bowley, procéder comme suit : on commence par ranger les classes dans un ordre régulier allant du minimum au maximum, en ayant soin d'observer que les classes aient la même grandeur : le contraire arrive fréquemment quand on utilise un matériel publié par un bureau de statistique.

En regard de chaque classe, on note les fréquences correspondantes et on examine la série ainsi formée. Si l'on a affaire avec une série régulière, obéissant à la loi normale des erreurs, il n'y aura guère de difficulté et la dominante se découvrira d'elle-même. Mais les séries régulières sont rares, comme nous l'avons déjà dit plusieurs fois et, dans la réalité, ce sont les séries asymétriques qui l'emportent : celles-ci ont plusieurs sommets et le tracé graphique qu'on en fait ressemble assez bien à une chaîne de montagnes, dont les pics, plus ou moins groupés autour

d'un massif central, sont séparés par de profondes vallées. Les séries irrégulières, peut-on dire encore, sont en dents de scie.

270. M. Bowley, dans l'exemple qui lui sert d'illustration, trouve 14 maxima dans la série ; dans le cas que nous citons ci-après, on peut compter 15 sommets différents, dont le plus élevé correspond au chiffre 1209.

Exemple 21.

Salaires de 10,455 ouvriers mâles de plus de 16 ans, en octobre 1903, dans l'industrie de la construction de machines motrices, machines outils, et appareils industriels, en Belgique (1).

CLASSE DES SALAIRES	Nombre d'ouvriers dans la classe ci-contre	CLASSE DES SALAIRES	Nombre d'ouvriers dans la classe ci contre	CLASSE DES SALAIRES	Nombre d'ouvriers dans la classe ci-contre
Fr.		Fr.		Fr.	
0.25 à 0.49	1	4.25 à 4 49	729	8 25 à 8.49	3
0.50 à 0.74	14	4.50 à 4.74	821	8.50 à 8.74	1
0.75 à 0.99	12	4.75 à 4.99	470	8.75 à 8.99	2
1.00 à 1.24	91	5.00 à 5.24	535	9 00 à 9.24	2
1.25 à 1.49	106	5.25 à 5.49	183	9.25 à 9.49	1
1.50 à 1.74	182	5.50 à 5.74	270	9.50 à 9 74	1
1.75 à 1.99	172	5.75 à 5.99	103	9.75 à 9.99	3
2.00 à 2.24	289	6.00 à 6.24	121	10 00 à 10.24	3
2.25 à 2.49	285	6.25 à 6.49	42	10.25 à 10.49	1
2.50 à 2.74	471	6.50 à 6.74	60	10.50 à 10.74	5
2.75 à 2.99	555	6.75 à 6.99	20	10.75 à 10.99	0
3.00 à 2.24	957	7.00 à 7.24	31	11.00 à 11.24	6
3.25 à 3.49	700	7.25 à 7.49	12	11.25 à 11.49	0
3.50 à 3 74	1029	7.50 à 7.74	16	11 50 à 11.74	0
3.75 à 3.99	926	7.75 à 7.99	10	11.75 à 11.99	1
4.00 à 4.24	1209	8.00 à 8.24	4		

(1) *Salaires et durée du travail dans les industries des métaux*, octobre 1903. Publication de l'Office du Travail de Belgique, Bruxelles, 1907.

271. Il s'agit de substituer à cette série extrêmement irrégulière, une série se rapprochant de la normale. Dans ce but, on procède comme suit : on débute par grouper deux à deux les fréquences, en commençant par la classe inscrite en tête du tableau et si l'on trouve plusieurs « modes », on continue le groupement par deux en commençant par la classe suivante (la seconde). On procède de même en groupant les classes trois par trois, en commençant par la première, puis par la seconde, enfin par la troisième. Si la régularité n'est pas encore obtenue, on continue en réunissant les classes quatre par quatre. Dans l'exemple donné par M. Bowley, il reste quatre « modes » quand on opère le premier et le second groupement, trois au troisième, quatre au quatrième, deux au cinquième, deux au sixième et seulement un au septième groupement. Si l'on s'arrête au sixième groupement, dont un des deux modes ne présente guère d'importance, on se trouve devant une classe fort étendue comprenant cinq classes du tableau primitif. M. Bowley propose, pour réduire l'étendue de cette classe, une méthode qu'on peut définir comme suit : augmenter d'une unité le nombre de classes à considérer et commencer le groupement par une classe située un degré plus bas que le point de départ du groupement précédent; continuer en opérant sur une base de même étendue et en descendant chaque fois d'un degré jusqu'à ce que la limite inférieure atteigne le degré supérieur de la classe qu'il s'agit de décomposer.

Le « mode » peut être fixé approximativement à la moitié de la classe comprenant la fréquence la plus élevée. Dans le cas cité par M. Bowley, cette fréquence étant 2012 et la classe correspondante étant celle des salaires de $ 1.05 à $ 1.55, le mode peut être placé *aux environs* de $ 1.25 à $ 1.34, c'est-à-dire $ 1.30.

M. Bowley signale une autre méthode approximative de déterminer le « mode », qui a de commun avec la première toutes les opérations préliminaires de groupement, mais

qui en diffère par le choix final. Lorsque le mode le plus important varie quand la limite la plus basse de la classe vient à changer, on peut en conclure que le groupement par classe est trop restreint et on peut passer à un groupement plus étendu. Lorsque, par exemple, on a déterminé les sommets au moyen du groupement par trois classes à la fois, on doit voir si l'une des classes est comprise dans les trois points déterminés à l'aide de ce groupement et le groupe le plus petit peut être pris comme contenant le mode.

272. Nous prenons comme exemple la série des salaires des ouvriers des ateliers de construction de machines motrices, machines-outils, etc., en octobre 1903, en Belgique, mais nous sommes tenté de trouver que la tâche est trop facile car dès le premier groupement par deux (0.25 à 0.49; 0.50 à 0.74, etc.) nous obtenons une série absolument régulière, dont la dominante fixée au point fr. 4.00 à 4.24 est confirmée par le second groupement par deux classes à la fois. Il n'est donc pas douteux, dès les premières variations, que la dominante soit contenue dans les limites fr. 4.00 à 4.24, soit fr. 4.12 environ.

La série I est : 15, 103, 288, 461, 756, 1512, 1729, 2135, 1550, 1005, 453, 224, etc.

La série II est : 26,197, 354, 574, 1026, 1657, 1955, 1938, 1291, 718, 373, 163, etc.

La classe des salaires de fr. 4.00 à 4.24 étant comprise dans les deux maxima, il est certain que le mode s'y trouve contenu. Aussi, pour avoir une série permettant de pousser l'analyse jusqu'au bout, avons-nous dû intervertir l'ordre de certaines fréquences, mais en prenant soin de ne modifier ni la médiane, ni la moyenne. La série hypothétique que nous avons ainsi construite comprend les données : 957, 926 (au lieu de 700), 1029, 700 (au lieu de 926), 729, 821, 470, 183 (au lieu de 535), 535 (au lieu de 183), etc. (le reste comme dans la série originale).

Le tableau suivant présente l'arrangement des données ainsi constituées.

Exemple 22.

Salaires de 10,455 ouvriers des ateliers de construction de machines motrices, etc., en Belgique. (Octobre 1903.)

Détermination de la dominante.

0.25 à 0.49 =	1	15					
0.50 à 0.74 =	**14**		26	27			118
0.75 à 0.99 =	12	103			117		
1.00 à 1.24 =	91		197			209	
1.25 à 1.49 =	106	288		379			
1 50 à 1 74 =	**182**		354		460		749
1.75 à 1.99 =	172	461				643	
2.00 à 2.24 =	**289**		**574**	746			
2 25 à 2.49 =	285	756			1045		
2.50 à 2.74 =	471		1026			1311	2268
2.75 à 2.99 =	555	1512		1983			
3.00 à 3.24 =	**957**		**1883**		2438		
3.25 à 3.49 =	926	**1955**				**2912**	
3.50 à 3.74 =	**1029**		1729	2655			3864
3.75 à 3.99 =	700	1909			**2938**		
4.00 à 4.24 =	**1209**		**1938**			2638	
4.25 à 4.49 =	729	1550		**2759**			
4.50 à 4.74 =	**821**		1291		2020		2203
4.75 à 4.99 =	470	653				1474	
5.00 à 5.24 =	183		718	1188			
5.25 à 5.49 =	**535**	**805**			988		
5.50 à 5.74 =	270		373			908	1029
5.75 à 5.99 =	103	224		494			
6 00 à 6.24 =	121		163		266		

6.25 à 6.49 =	42	102				223	
6.50 à 6.74 =	**60**		80	122			153
6.75 à 6.99 =	20	51			111		
7.00 à 7.24 =	**31**		43			63	
7.25 à 7.49 =	12	28		59			
7.50 à 7.74 =	**16**		26		38		42
7.75 à 7.99 =	10	14				30	
8.00 à 8.24 =	4		7	17			
8.25 à 8.49 =	3	4			8		
8.50 à 8.74 =	1		3			6	8
8.75 à 8.99 =	2	4		5			
9.00 à 9.24 =	2		3		5		
9.25 à 9.49 =	1	2				4	
9.50 à 9.74 =	1		4	5			8
9.75 à 9.99 =	3	6			7		
10.00 à 10.24 =	3		4			7	
10.25 à 10.49 =	1	6		9			
10.50 à 10.74 =	5		5		6		12
10.75 à 10.99 =	0	6				11	
11.00 à 11.24 =	**6**		6	6			
11.25 à 11.49 =	0	0			6		
11.50 à 11.74 =	0		1	1		1	
11.75 à 11.99 =	1	1			1		1

La série originaire comprend 13 sommets dont le plus typique est 1209. En sériant une première fois les données deux à deux, il reste deux sommets : le second groupement par deux en donne aussi deux qui ne coïncident pas

avec les premiers, de telle sorte que nous sommes forcé de recourir à un groupement plus étendu. En groupant par trois, nous n'avons plus qu'un mode, mais ce mode change chaque fois qu'on descend d'un degré en commençant la sériation. M. Bowley pose alors la règle suivante : entre les trois modes, choisir celui qui est le plus petit et fixer le mode à peu près au milieu de la classe centrale ; dans notre exemple, ce mode est 2759, la classe fr. 4.25 à fr. 4.49 et l'emplacement du mode 4.37 environ. Si donc on demande quel est le salaire le plus habituel parmi les ouvriers belges des ateliers de construction de machines, on pourrait répondre que ce salaire varie de fr. 4.25 à 4.49.

273. La méthode de M. Bowley est assez peu précise ; nous voyons par l'exemple même qu'il donne que le mode varie d'une classe selon qu'on emploie la première ou la seconde méthode. M. Pearson a donné une autre méthode approximative qu'on peut résumer de la sorte :

1° Calculer la moyenne de la série ;

2° Calculer la médiane de la série ;

3° On a le mode = la moyenne — 3 (moyenne — médiane), c'est-à-dire que le mode se trouve placé du côté de la médiane, à une distance de celle-ci qui équivaut au double de la distance entre la médiane et la moyenne, ou que la distance entre la moyenne et la médiane égale le tiers de la distance entre la moyenne et le mode.

Soit l'exemple suivant (1) :

Pourcentage de la population secourue en Angleterre et dans le pays de Galles :

---

(1) Yule, « Notes on the history of pauperism; supplementary note on the determination of the mode ». (*Journal of the Roy. Stat. Soc.*, 1896, p. 343.)

Année 1891, moyenne = 3,289
médiane = 3,195

différence = 0,094

3,289 — (3 × 0,094) = 3,007 (1)
car 0,094 × 3 = 0,282
et 3,289 — 0,282 = 3,007

Le mode « idéal » serait : 2,987 ou 2,99.

274. Si la dominante réunit de nombreux avantages, elle présente aussi quelques inconvénients, parmi lesquels le principal est de ne pouvoir être calculée avec une précision véritable par des procédés simples. Les mesures qui en sont données sont souvent basées sur une recherche empirique n'aboutissant qu'à un résultat approximatif. Au contraire, la moyenne peut être connue avec la dernière précision en employant des procédés élémentaires. La recommandation que fait M. Yule à cet égard est significative : « La moyenne arithmétique, dit-il, doit être employée invariablement, à moins qu'il n'existe une raison très précise en faveur du choix d'une autre forme et le commençant fera bien de se limiter à la recherche de la moyenne arithmétique. » Et il ajoute : « S'il y a de bonnes raisons pour rechercher le « mode » en outre de la moyenne, il n'en est pas moins vrai que le « mode » ne peut remplacer la moyenne. » La dominante enfin renseigne une situation de fait, mais elle n'implique aucune idée de partage égal comme c'est le cas pour la moyenne. Si nous calculons le revenu moyen d'un groupe bien défini, la valeur moyenne nous fait connaître quel serait le revenu de chacun si la distribution des revenus était faite sur une base uniforme : le mode ne ferait que nous indiquer le revenu le plus usuel parmi les membres de la communauté dont il s'agit. Il

(1) G. U. Yule, *An introduction to the theory of statistics*, p. 121.

importe donc de bien poser le problème et d'examiner avant tout la nature exacte de la réponse qu'on désire obtenir. De ce qui précède, se dégage encore le corollaire suivant : en multipliant la moyenne par le nombre d'unités, on reconstitue le total, ce qui n'est pas exact si, au lieu de la moyenne, on emploie la dominante.

275. *Références.*

BERTILLON (J.), *Cours élémentaire de statistique administrative.* Paris, 1895, ch. IX, p. 100 et suivantes.

BERTILLON (A.), « La théorie des moyennes en statistique ». (*Journal de la Société de statistique de Paris,* 17e année, 1876, p. 266.)

ID., Cfr. article « Moyennes » dans le *Dictionnaire encyclopédique des sciences médicales.*

BLOCK (M.), *Traité théorique et pratique de statistique.* Paris, 1886, p. 121 et suivantes.

BOSCO (A.), *Lezioni di statistica,* parte prima. Roma 1909, p. 445 et suiv. Pour la médiane, *eod. loc.,* pp. 493-497 ; pour le mode, pp. 479-492.

BOWLEY (A. L.), *Elements of statistics.* London, 1901, p. 107 et suivantes. Pour la médiane, *eod. loc.,* pp. 125-128 ; pour le mode, pp. 118-124.

COLAJANNI (N.), *Manuale di statistica teorica,* 3e édit. Napoli, 1910, p. 182 et suivantes.

CRAWFORD (G.), « An Elementary Proof that the arithmetic mean of any number of positive quantities is greater than the geometric mean ». (*Proceedings of the Edimb. math. Soc.,* vol. XVIII, 1899-1900.)

DEWEY (John), *Galton's statistical methods* (Quarterly publications of the American statistical association, new series, n° 8, 1888).

EDGEWORTH (F. Y.), « On the method of ascertaining a change in the value of gold ». (*Journal of the Statist. soc.,* 1883, p. 714.)

ID., « On methods of statistics ». (*Jubilee volume of the Statist. soc.* (London, 1885, p. 181.)

ID., « Some news methods of measuring variations in general prices ». (*Journal of the Roy. Statist. Soc.* London, 1888, p. 346 et suivantes.)

ELDERTON (Palin and Ethel), *Primer of statistics.* London, 1912 ; pour la médiane, cfr. pp. 1-22.

FECHNER, *Ueber den Ausgangswert der kleinsten Abweichungssummen* (XIe Band der abt. der mathem. phys. Klasse der K. Sachs. Gesellschaft der Wissench, n° 1. Leipzig, 1874).

FLUX (Prof.), « Modes of constructing Index numbers ». (*Quart. Journal of Economics,* vol. XXI, p. 613.)

GABAGLIO (A.), *Teoria generale della statistica*, vol. II. Milano, 1888, p. 201 et suivantes.

GALTON (Sir Francis), « The geometric mean in the vital and social statistics ». (*Proceedings of the Roy. Soc.*, vol. XXIX, 1879, p. 365.)

ID., « Application of the method of Percentiles to Mr. Yule's data on the distribution of pauperism ». (*Journal Roy. Statist. Soc.*, 1896.)

ID., *Natural inheritance*. London, 1889, p. 35 et suivantes.

HOLMES (G. K.), « A plea for the average ». (*Quart. public. of the Am. Statist. Assoc.*, new series, n° 16, décembre 1891.)

JEVONS (Stanley), « On the variation of prices and the value of currency since 1782 ». (*Journal of the Roy. Statist. Soc.*, vol. XXVIII, 1865; réimprimé dans *Investigations in currency and finance*. London, 1884.)

ID., « A serious Fall in the value of gold ascertained and its social effects ». London 1863. (Réimprimé dans *Investigations in currency and finance*, London, 1884.)

ID., *The principles of science, a treatise on logic and scientific method*. London, 1874, ch XVI, p. 357 et suivantes.

JULIN (A.), *Précis du cours de statistique*, 4e édition. Bruxelles, 1919, p. 69 et suivantes.

KING (W. J.), *The elements of statistical method*. New York, 1912, p. 121 et suivantes; pour la médiane, cfr. *eod. loc.*, pp. 127-132, et pour le mode, pp. 122-127.

LIESSE (A.), *La statistique, etc.* Paris, 1905, p. 73 et suivantes.

MAC-ALISTER, « On the use of the geometric mean in statistics ». (*Proceedings of the Roy. Statist. Soc.*, vol. XXIX, p. 367.)

MARCH (L.), « Essai sur un mode d'exposer les principaux éléments de la théorie statistique ». (*Journal de la Soc. de statist. de Paris*, 1910, p. 447 et suivantes.)

ID., « Quelques exemples de distribution des salaires ». (*Journal de la Soc. de statist. de Paris*, 1898, p. 201.)

MESSEDAGLIA (A.), « Calcul des valeurs moyennes ». (*Annales de démographie internationale*, t. IV, 1880, p. 387 et suivantes.) Le même en italien dans l'*Archivio di statistica*, anno V, 1880.

PEARSON (K.), « Skew variations in homogeneous material ». (*Phil. Trans. Roy. Soc.*, series A, vol. CLXXXVI, 1895, note p. 345.)

ID., *On the modal value of an organ or character*, *Biometrika*, vol. I, 1902, p. 260.

QUETELET (A.), « Sur l'appréciation des moyennes ». (*Bulletin de la Commission centrale de statistique*, t. II, 1845.)

ID., *Lettres sur la théorie des probabilités*. Bruxelles, 1846.

ID., *Physique sociale*. Bruxelles, 1869, vol. I, p. 486.

ID., *Anthropométrie ou mesure des différentes facultés de l'homme*. Bruxelles, 1871, n° 18.

VENN (J.), « On the nature and uses of averages ». (*Journal of the Roy. Statist. Soc.*, 1891, p. 36 et suivantes.)

VERRYN-STUART, *Inleiding tot de beoefening der statistiek,* 1ste deel Harlem, 1910, p. 35 et suivantes.

VIRGILII (F.), *Statistica,* 5e édition. Milano, 1911, p. 77 et suivantes.

YULE (U. G.), *An Introduction to the theory of statistics.* London, 1912 p. 106 et suivantes: pour la médiane, cfr. pp. 116-120; pour le mode, pp. 120-123.

ID., « Notes on the history of pauperism. Supplementary note on the determination of the mode ». (*Journal of the Roy. Statist. Soc.*, t. LIX, 1896, p. 346 et suivantes.)

ZIZEK, *Statistical averages,* trad. anglaise. New-York, 1913, p. 92 et suiv. Pour la médiane, cfr. pp. 199-221; pour le mode, pp. 222-247.

CHAPITRE III

La dispersion et ses mesures

I. — La nature de la dispersion.

276. Par dispersion, on entend, d'une manière générale, le manque d'uniformité des termes de la série, par rapport à la moyenne ou à une autre expression générale de la série. Lorsqu'une série se compose d'unités classées d'après leur grandeur, comme des soldats rangés par degré de taille, des ouvriers groupés selon le taux de leur salaire, on observe que les termes extrêmes sont dans les différents cas plus ou moins différents de la moyenne. Si tous les termes de la série étaient égaux, ou à peu près, ils prendraient tous place dans la même classe et il n'y aurait plus de dispersion du tout; s'ils ne différaient pas beaucoup, il ne faudrait que quelques classes, voisines l'une de l'autre, pour les contenir tous et la dispersion serait restreinte. Ainsi, la série des ouvriers adultes employés dans l'industrie de la construction de machines motrices, machines-outils et appareils industriels (Cfr. ex. 21, n° 270) présente une grande dispersion; par contre, dans l'exemple des 94 amandes (Cfr. ex. 18, n° 259), la dispersion est restreinte ainsi que le montre le tableau ci-après :

Exemple 1.

Dimensions des amandes — Centimètres.	Nombre d'amandes. —
2.50 à 2.74	8
2.75 à 2.99	14
3.00 à 3.24	24
3.25 à 3.49	24
3.50 à 3.74	18
3.75 à 3.99	5
4.00	1

On conçoit l'utilité de déterminer une mesure exacte de la dispersion, tant pour caractériser la nature d'une série particulière que pour comparer plusieurs séries différentes. Nous restons donc exactement dans la matière de ce livre second, qui a trait aux procédés d'analyse du matériel statistique.

277. Avant de passer aux procédés de calcul, il est utile de se bien pénétrer de la notion de la dispersion.

Les hommes d'une compagnie tirent sur une cible ; on relève la trace des balles qui ont porté ; bien que tous aient visé le centre de la cible, ce but n'a pas été atteint par tous les soldats — les uns s'en sont écartés de peu, les autres de beaucoup. En notant les coups, on obtient l'expression de la dispersion des balles dans la cible.

Des ouvriers sont employés dans une industrie ; supposons que ce soit l'industrie de la construction des machines d'un pays déterminé, à une date fixée. Les uns gagnent un faible salaire, d'autres un salaire élevé, le plus grand nombre un salaire intermédiaire entre les extrêmes. La répartition des ouvriers d'après le taux de salaire fournit le matériel nécessaire à l'étude de la dispersion des salaires dans cette industrie.

On dit que la dispersion est étendue lorsque les termes de la série s'écartent beaucoup de la grandeur moyenne. Elle est restreinte lorsque les différences avec la moyenne sont atténuées.

Il est clair que le résultat du tir ou le fait de gagner un salaire déterminé n'est pas l'effet du hasard. Chaque tireur, stimulé par l'émulation qui règne entre les concurrents, s'efforce de placer sa balle au centre ; une circonstance extrinsèque joue un rôle important : la justesse de l'arme, son réglage plus ou moins complet ; il faut ensuite

faire la part des conditions intrinsèques telles que le coup d'œil, le calme, la sûreté du tireur, etc., bref ce qui caractérise le tireur adroit. La réalisation plus ou moins parfaite de ces conditions, ou leur absence, se traduisent finalement dans le résultat du tir. Ce résultat est plus ou moins bon selon que les circonstances agissantes ont formé un ensemble favorable ou non. Des balles tirées très loin du centre marqueront qu'il y a dans la compagnie quelques soldats fort nerveux ou peu expérimentés (si l'on a tous ses apaisements sur la valeur des armes). Un grand nombre de balles avoisinant le centre fera supposer qu'il y a une bonne moyenne de tireurs au courant de cet art, et un autre groupe situé en dehors des premiers cercles de la cible donnera à penser qu'il existe encore pas mal de novices. Si l'on peut caractériser, par certains procédés de calcul, la dispersion des termes de la série, on aura rendu un grand service aux chercheurs et on aura donné à l'homme d'étude l'expression synthétique dont il a besoin. Telle est l'utilité des mesures de dispersion, dans son expression la plus générale.

278. La notion de dispersion apparaît plus claire encore lorsqu'on choisit un exemple emprunté à la vie sociale, par exemple la distribution des salaires dans un groupe déterminé d'ouvriers. Reprenons les chiffres donnés à notre exemple 21 (chap. II, n° 270), mais, pour simplifier, donnons les variations des salaires de cinquante en cinquante centimes, tels qu'ils sont déterminés à l'exemple 22, 3e colonne.

EXEMPLE 2. — **Salaires de 10,455 ouvriers des ateliers de construction de machines motrices, etc , en Belgique, octobre 1903.** (Matériel extrait de la statistique des salaires dans les industries des métaux, publication de l'Office du Travail de Belgique.)

Classes de salaires	Nombre d'ouvriers	Classes de salaires	Nombre d'ouvriers
0 25 à 0.74	15	6.25 à 6.74	102
0.75 à 1.24	103	6.75 à 7.24	51
1.25 à 1.74	288	7.25 à 7.74	28
1.75 à 2.24	461	7.75 à 8.24	14
2.25 à 2.74	756	8.25 à 8.74	4
2.75 à 3.24	1512	8.75 à 9.24	4
3.25 à 3.74	1955	9.25 à 9.74	2
3.75 à 4.24	1909	9.75 à 10.24	6
4.25 à 4.74	1550	10 25 à 10.74	6
4.75 à 5 24	653	10.75 à 11.24	6
5.25 à 5.74	805	11.25 à 11.74	C
5.75 à 6.24	224	11.75 à 11.99	1

Cette distribution des salaires d'un groupe important d'ouvriers confirme les observations que M. Lucien March présentait (1), à propos des courbes de distribution des salaires, à savoir : 1° les salaires sont concentrés autour d'une valeur normale ; il y a donc une tendance au maintien du taux du salaire d'un type déterminé ; 2° à mesure que les salaires augmentent, leur dispersion à partir de la valeur normale et la dissymétrie de la courbe de distribution augmentent également. L'accroissement de la dispersion ne s'opère pas d'une manière symétrique. La courbe de distribution se déforme, non pas comme si toute la masse des ouvriers participait en même temps à la hausse des salaires, mais comme s'il y avait un effort, une tendance con-

(1) MARCH (L.), « Quelques exemples de distribution des salaires ». (*Journal de la Société de statistique de Paris,* 1898, pp. 202-203.)

stante, venant des ouvriers à salaires élevés, en sorte que l'écart entre la moyenne et la normale tend à s'accroître constamment.

Il paraît inutile, après ces exemples, d'insister sur l'utilité d'une formule représentative de la dispersion des phénomènes sociaux.

279. L'idée qui se présente à première vue pour donner une expression synthétique de la dispersion est de considérer l'écart entre les termes extrêmes rapporté à la grandeur moyenne de la série. Ce procédé doit être repoussé pour une raison bien simple; c'est qu'il ne s'agit pas uniquement de classes, mais aussi de fréquences dans chaque classe. Ainsi dans l'exemple précédent, la dispersion se trouverait notablement moindre si l'on supprimait seulement 25 ouvriers sur le total de 10,455; elle s'arrêterait, en effet, à la limite fr. 8.74, au lieu de comprendre encore 7 classes de fr. 0.50 pour aller jusqu'à fr. 11.99. L'addition de quelques cas exceptionnels représentant $\frac{25}{10455}$ suffirait donc à augmenter la grandeur de la dispersion dans de fortes proportions.

Aussi les mesures adoptées pour caractériser la dispersion d'une série sont-elles empruntées à d'autres éléments. Nous les examinerons successivement. Elle peuvent se ramener à trois types qui sont :

La moyenne de la déviation (1);

La déviation type;

La méthode des quartiles.

On peut y ajouter :

Le coefficient de variation et la dissymétrie ou « Skewness ».

---

(1) On pourrait dire aussi la moyenne des écarts, mais cette appellation prêterait à confusion, parce que la somme des écarts, avec leurs signes, étant égale à zéro, leur moyenne serait aussi égale à zéro. Le mot déviation évite ce malentendu et il présente l'avantage de respecter la terminologie employée par ceux qui ont introduit l'expression en statistique.

II. — Moyenne de déviation.

280. Des différentes mesures de la dispersion, la moyenne de déviation est la plus simple et la plus rapidement calculée. On a vu plus haut ce qu'on entend par « écarts » et l'on a démontré que la somme algébrique des écarts à la moyenne arithmétique est égale à zéro. Mais, si au lieu de donner aux écarts leurs signes propres, on les considère tous comme étant de signe positif, on obtient alors une valeur positive, qui, divisée par le nombre de termes, prend le nom de moyenne de déviation (mean deviation).

Cette moyenne de déviation, désignée habituellement par la lettre Δ peut se calculer d'après des écarts pris par rapport à la moyenne ou à toute autre mesure. Cependant, il est recommandé de calculer les écarts des nombres à la médiane, de préférence à toute autre mesure. En effet, comme Laplace l'a démontré en 1818 déjà, la moyenne de la déviation est la plus petite possible quand on utilise la médiane comme point de comparaison.

Pour calculer la moyenne de déviation d'une série composée de nombres sans distribution de grandeur, il n'y a qu'à se rappeler les règles de calcul de la médiane.

Nous avons dit plus haut (Cfr. n° 256) que si le nombre de variables est impair, il suffit de laisser tomber à droite et à gauche un nombre égal de variables et d'attribuer la valeur médiane à la grandeur, représentée par une seule variable, placée entre les deux groupes ainsi constitués, les grandeurs étant placées dans leur ordre naturel.

La moyenne de déviation s'exprime au moyen de la formule suivante :

$$\delta_{\text{Mé}} = \Sigma \frac{(m - \text{Mé})}{n} \text{ ou } \frac{\Sigma d^{\text{Mé}}}{n} \qquad (27)$$

Soient les nombres, déjà utilisés pour des exemples antérieurs : 19, 25, 26, 28, 32, 33, 40. Par la simple inspection

des nombres, nous voyons que la médiane est 28. Nous avons donc :

EXEMPLE 3.

19	. . . . . .	x_1	9
25	. . . . . .	x_2	3
26	. . . . . .	x_3	2
28	. . . . . .	x_4	0
32	. . . . . .	x_5	4
33	. . . . . .	x_6	5
40	. . . . . .	x_7	12
		x_n	35

$$\Delta = \frac{35}{7} = 5$$

La moyenne arithmétique des mêmes nombres est 29. Calculant la moyenne de déviation en adoptant cette base de la moyenne, ainsi qu'on l'a fait à l'exemple précédent, nous avons :

EXEMPLE 4.

19	. . . . . .	x_1	10
25	. . . . . .	x_2	4
26	. . . . . .	x_3	3
28	. . . . . .	x_4	1
32	. . . . . .	x_5	3
33	. . . . . .	x_6	4
40	. . . . . .	x_7	11
		x_n	36

$$\Delta = \frac{36}{7} = 5{,}142$$

La moyenne de déviation est donc ici trouvée plus petite quand on la calcule d'après la médiane que quand on la prend par rapport à la moyenne, résultat conforme à la théorie (1).

(1) La moyenne de déviation a un rapport direct, comme on le verra plus loin, avec la *standard deviation*, ou déviation-type. Il existe un moyen rapide de contrôler, l'un par l'autre, les résultats du calcul : en effet, la moyenne de déviation représente environ les 4/5 de la *standard deviation*. Le rapport entre les deux valeurs est donc : $\frac{\text{moyenne de déviation (d'après médiane)}}{\text{déviation type.}}$

On reviendra sur ce point au n° 286.

281. Donnons encore un exemple d'après une série composée de nombres se succédant l'un l'autre. D'après le tableau inséré au n° 259, nous avons vu que, pour 94 amandes, dont la longueur avait été soigneusement mesurée, nous avons trouvé que les écarts à la médiane étaient :

1.260 + 1.258. Nous avons donc :

$$\Sigma\, x_n = \frac{1.260 + 1.258}{94} = 0.0267$$

moyenne de déviation.

La règle précédente est excessivement simple. Elle n'est pas plus compliquée quand la médiane doit être calculée sur une série de nombres pairs : nous avons indiqué la manière de procéder à la détermination de la médiane dans un tel cas (Cfr. n° 256). Rappelons qu'on peut procéder, soit en prenant la moyenne des deux variables formant les extrémités médianes des deux groupes qui partagent la série, soit en ayant recours à un procédé d'interpolation.

282. Mais un grand nombre de séries statistiques sont composées de variables ayant des fréquences différentes, dans ce cas, comment convient-il de procéder ?

Il paraît naturel, à première vue, d'établir les calculs comme on le ferait s'il s'agissait de déterminer simplement les écarts à la moyenne arithmétique ; cependant, ce procédé ne conduirait à aucun résultat. La méthode à employer est celle que nous avons appelée méthode indirecte (Cfr. n° 230) dans laquelle l'éloignement de la classe de chaque variable à la moyenne est la mesure de l'importance des écarts.

On peut donc formuler la règle d'une manière très simple : *la moyenne de déviation est la somme des déviations de chaque classe multipliées par la fréquence de la classe, somme qui est divisée par le nombre de variables* (1).

(1) DAVENPORT, *Statistical methods*, p. 16.

Appliquons la règle à la recherche de la moyenne de déviation du *Pimpinella Saxifraga L.* d'après la médiane; celle-ci tombe dans la classe 11.

EXEMPLE 5.

Classes	Fréquences (f)	Écarts (ξ)	Produits ($f\xi$)
5	1	6	6
6	5	5	25
7	9	4	36
8	22	3	66
9	38	2	76
10	62	1	62
11	61	0	0
12	29	1	29
13	14	2	28
14	4	3	12
15	4	4	16
	249		$\Sigma(f\xi)$ 356

$$\Delta = \frac{356}{249} = 1.429$$

283. Cependant, pour arriver à un résultat d'une exactitude parfaite, il faut tenir compte de l'influence exercée par l'intervalle de classe. La manière de procéder est alors la suivante : on fait la somme des produits des déviations par les fréquences; ensuite on calcule le nombre de fréquences au-dessous de la moyenne et le nombre de fréquences restant; on soustrait ces nombres l'un de l'autre et on multiplie le reste par l'intervalle de classe; le produit est soustrait de Σ (f. ξ) et le reste est divisé par le nombre de variables.

Appliquons la règle aux données de l'exemple 8 du chapitre précédent (Cfr. n° 230). Le nombre total de variables est 433, la somme Σ (f. δ) $= 1458$; appelons V_1 les varia-

bles au-dessous de la moyenne et V_2 les autres : $V_1 - V_2 = -571+887=+316$. L'intervalle de classe $= \frac{316}{433} / 4 = 0{,}182$. Nous avons : $316 \times 0.182 = 57.512$;

$$\Delta = \frac{1458 - 57.512}{433} = 3.234$$

Les avantages de la moyenne de déviation sont faciles à discerner. La notion de l'écart des termes à la médiane ou à la moyenne est des plus simples, même en tenant compte de l'intervalle de classe et on ne doit faire appel, dans ce calcul, qu'aux procédés arithmétiques élémentaires. Dans le cas d'une série groupée, les opérations consécutives se bornent à des multiplications et divisions rapidement effectuées. Le calcul n'offre ainsi aucune complication, ce qui est toujours un précieux avantage. La moyenne de déviation a de nombreuses applications aux questions économiques, principalement à celles qui concernent la distribution de la richesse parmi les citoyens d'un Etat.

III. — La déviation-type.

284. (Standard deviation). La définition de la standard deviation a été donnée en 1894 par M. Pearson dans son premier mémoire faisant partie de l'ensemble intitulé : « Contributions to the mathematical theory of evolution. » Cette mesure est *la racine carrée de la moyenne arithmétique des carrés des écarts.* L'expression standard déviation est le terme sous lequel on désigne cette mesure de variabilité, mais en calcul des probabilités elle porte le nom de « erreur moyenne quadratique ». On l'appelle aussi « erreur moyenne » ou « erreur moyenne à craindre » (Gauss), mais pour éviter les confusions il est préférable de réserver le nom d'erreur moyenne à l'erreur première représentée par

$$e_1 = \Sigma \frac{x}{N}$$

et d'appeler erreur carrée ou quadratique moyenne celle de la forme

$$e_2 = \sqrt[2]{\Sigma \frac{x^2}{N}}$$

Comme aussi on peut parler de l'erreur cubique moyenne

$$e_3 = \sqrt[3]{\Sigma \frac{x^3}{N}}$$

Ces diverses mesures présentent entre elles des relations très curieuses ; par exemple, on démontre que

$$2 \left(\frac{e_2}{e_1}\right)^2 = \pi$$

lorsqu'on a un très grand nombre d'observations, toutes d'égale précision (1).

La déviation-type ou standard déviation est représentée par la lettre σ et sa formule est la suivante, donnée par l'équation :

$$\sigma = \sqrt[2]{\frac{1}{N} \Sigma (x^2)} \qquad (28^{A})$$

Il est important de rappeler ici que, d'après la seconde propriété des moyennes, la somme des carrés des écarts à la moyenne est un minimum (Cfr. n° 242).

La formule, telle qu'elle est écrite plus haut, est très claire et se rapproche sensiblement de la formule de la moyenne arithmétique simple.

La formule de la déviation type ou standard déviation peut aussi s'écrire :

$$\sqrt{\left\{\frac{fa\,a'^2 + fb\,b'^2 + \ldots\ldots + fn\,n'^2}{fa + fb + \ldots\ldots + fn}\right\}} \qquad (28^{B}) \qquad (2)$$

(1) Boudin-Mansion, *loc. cit.*, pp. 174 et 239.

(2) Elderton (Palin W.), *Frequency-curves and correlations.* Londres, Layton, 1906, p. 10.

fa, fb, fn désignent les fréquences et *a', b', n'*, l'écart par rapport de la moyenne.

La grandeur de la standard déviation montre quel est l'espacement des fréquences autour de la moyenne : si la déviation-type s'exprime par une fraction proche de l'unité, c'est que la dispersion autour de la moyenne est considérable; au contraire, la fraction diminue d'autant plus que les fréquences sont resserrées autour de la valeur centrale.

Dans les formules ci-dessus, on a en vue le cas le plus simple, celui où une succession de nombres, ayant tous un poids égal, constitue la série, comme dans l'exemple suivant :

Exemple 6. — **Nombre de mariages en Angleterre et dans le Pays de Galles de 1891 à 1900**
(M = 239,410.5)

ANNÉES	Nombre de mariages	Écarts à la moyenne	x^2
1891	226,526	— 12,884.5	166,010,340.25
1892	227,135	— 12,275.5	150,687,900.25
1893	218,689	— 20,721.5	429,380,562.25
1894	226,449	— 12,961.5	168,000,482.25
1895	228,204	— 11,206.5	125,585,642.25
1896	242,764	3,353.5	11,242,609.25
1897	249,145	9,734.5	94,750,756.25
1898	255,379	15,968.5	254,992,992.25
1899	262 334	22,923.5	525,486,852.25
1900	257,480	18,069.5	326,506,830.25
			2,252,644,967.50

$$\sqrt{\frac{2,252,644,967.50}{10}} =$$

$$= \sqrt{225,264,496} =$$

$$= 15,009$$

285. D'après ce qui précède et en se souvenant de ce qui a été exposé à propos du procédé indirect de recherche de la moyenne (Cfr. n° 229) il n'est pas difficile de saisir une formule générale d'une expression analogue à la standard déviation, basée sur une valeur autre que M.

Lorsque nous prenons pour origine une valeur autre que la moyenne arithmétique et que nous calculons l'écart (ξ) de chaque terme (X) à cette valeur arbitraire (A), nous avons (1) :

$$\xi = X - A$$

et la formule de l'expression cherchée est analogue à celle de la standard déviation (form. 28^A) ; nous écrivons (s^2 étant le symbole de l'expression nouvelle) :

$$s^2 = \frac{1}{N} \Sigma (\xi^2) \qquad (29)$$

Quant au rapport de s^2 à σ^2, il est aisé à définir :

$$s^2 = \sigma^2$$

plus l'écart entre la moyenne véritable et la valeur arbitraire choisie, que nous appellerons d :

$$s^2 = \sigma^2 + d^2 \qquad (30)$$

Faisons application de la règle à l'exemple 4 du chapitre II. Il reste entendu que l'intervalle de chaque classe est partagé en deux parties égales $\frac{(0.25)}{2}$ de sorte qu'au taux initial il suffit d'ajouter chaque fois 0.125. De même que dans l'exemple 7 du chapitre II, au lieu de la moyenne 3.807, prenons comme valeur arbitraire (A), origine des

(1) Pour cette démonstration et ce qui suit, cfr. l'exposé détaillé de Yule. *ouvrage cité*, p. 134 et suivantes.

écarts, le point 3.625. Nous procédons alors comme suit : du point d'origine, ayant pour valeur 0 nous remontons vers les taux les plus bas en attribuant successivement à chaque classe les valeurs 1, 2, 3, etc., et nous procédons de même, en sens inverse, en partant de l'origine 0, en descendant vers les taux les plus élevés. Nous faisons ensuite le produit de ces valeurs 1, 2, 3, etc., par les fréquences inscrites en regard des classes et enfin le produit des mêmes fréquences par le carré des écarts ou, pour procéder d'une manière plus rapide, nous multiplions encore une fois le produit obtenu par les fréquences, puis nous additionnons les produits. Nous rappelant d'après l'exemple 8 du chapitre II, que l'intervalle de la classe est 0.182, nous divisons le dernier produit total de la colonne 5 par le nombre total des fréquences, et nous soustrayons du quotient la valeur 0.182 au carré; enfin, nous extrayons la racine carrée du reste et nous tenons compte de la grandeur de l'intervalle de la classe. (*Voir tableau ci-contre.*)

EXEMPLE 7

CLASSES	Fréquences (f.)	Deviation par rapport à A = 3.625 ξ)	Produits (f. ξ)	Produits (f. ξ)²
Fr.				
1.50 à 1.74	17	— 8	136	1088
1.75 à 1.99	3	— 7	21	147
2.00 à 2.24	10	— 6	60	360
2.25 à 2 49	8	— 5	40	200
2 50 à 2.74	34	— 4	136	544
2.75 à 2.99	28	— 3	84	252
3.00 à 3.24	34	— 2	68	136
3.25 à 3.49	26	— 1	26	26
3.50 à 3.74	47	0	— 571	—
3.75 à 3.99	44	+ 1	44	44
4.00 à 4.24	60	+ 2	120	240
4.25 à 4.49	24	+ 3	72	216
4.50 à 4.74	27	+ 4	108	432
4.75 à 4 99	9	+ 5	45	225
5.00 à 5.24	22	+ 6	132	792
5 25 à 5.49	8	+ 7	56	392
5.50 à 5.74	5	+ 8	40	320
5.75 à 5.99	13	+ 9	117	1053
6.00 à 6.24	8	+ 10	80	800
6.25 à 6.49	2	+ 11	22	242
6.50 à 6.74	1	+ 12	12	144
6.75 à 6.99	3	+ 13	39	507
	433		+ 887	8160

$$\frac{8160}{433} = 18.845$$

$$(0.182)^2 = 0.033$$

$$\sigma^2 = \sqrt{18.812}$$

$$\sigma = 4.337$$

286. Une règle de calcul un peu différente peut être formulée en ces termes (1) : adopter pour moyenne une certaine valeur arbitraire proche de la moyenne véritable : établir les écarts de chaque terme à cette valeur : en faire le carré, faire la somme des carrés : soustraire n fois le carré de la différence entre le nombre assumé et la moyenne vraie : diviser par n : extraire la racine carrée du quotient.

La formule exprimant cette méthode est la suivante :

$$\sigma = \sqrt{\frac{\Sigma(m - x)^2 - n(a - x)^2}{n}}$$

Reprenons l'exemple précédent et traitons-le d'après la règle qui vient d'être énoncée.

Nous savons, en premier lieu, que M = 3.807.

La valeur arbitraire assumée est A = 3.625.

$$M - A = -0.182$$
$$(M - A)^2 = 0{,}03312$$
$$n\,(M - A)^2 = -14.340$$
$$\text{Le produit } \Sigma\,(f\xi)^2 = 8160$$

$$\sqrt{\frac{8160 - 14.340}{433}} = \frac{8145.660}{433} = \sqrt{18.806} = 4.337$$

résultat identique à celui obtenu précédemment.

Nous avons vu (Cfr. n° 250) qu'on démontre facilement que la moyenne d'une série entière est égale à la somme des moyennes particulières, divisée par le nombre de ces moyennes, formées à l'aide des termes de la série, si ces termes ont un poids égal. M. Yule démontre la même propriété en ce qui concerne la standard-déviation (2). Appe-

---

(1) D'après KING (W. I.), *Elements of statistical method*, New-York, 1912, p. 149.

Cfr. *An Introduction to the theory of statistics*, p. 142.

lons M la moyenne d'une série entière, M_1 et M_2 les moyennes de parties composant cette série, d_1 et d_2 la différence, nous avons évidemment

$$M_1 - M = d_1$$
$$M_2 - M = d_2$$

et les standard déviations à partir de M sont, en raison de ce qui a été dit plus haut,

$$\sigma_1^2 + d_1^2$$
$$\sigma_2^2 + d_2^2$$

Donc

$$N.\ \sigma^2 = N_1\ (\sigma_1^2 + d_1^2) + N_2\ (\sigma_2^2 + d_2^2)$$

et par conséquent, en admettant que les termes des deux séries particulières soient en nombre égal $\left(N_1 \mp N_2 = \frac{N}{2}\right)$

$$\sigma^2 = \frac{(\sigma_1^2 + \sigma_2^2)}{2}$$

287. Ainsi que nous l'avons déjà fait observer (Cfr. n° 280, note) il existe entre la moyenne de déviation et la déviation-type un rapport constant qui peut se traduire par environ 0.80 (0.7979) (1), cette mesure marquant le rapport de la mean-déviation à la standard-déviation. Cette mesure n'est qu'approximative, mais comme le fait remarquer M. Yule, elle présente au moins l'avantage de vérifier l'exactitude du calcul arithmétique et d'empêcher les erreurs les plus grossières. Ainsi, par exemple, nous voyons que pour les nombres 19, 25, 28, 32, 33 et 40, $\Delta = 5$; la déviation type pour les mêmes données est :

$$\frac{\Sigma\ x\ n}{n} = \frac{272}{7} = \sqrt{38.857} = 6.233$$

or :

$$\frac{5.000}{6\ 233} = 0.802$$

(1) Davenport, *Statistical methods*, p. 16.

Une des deux mesures étant connue, on obtient l'autre par une simple opération arithmétique. Son rôle essentiel consiste à faciliter la vérification de calculs souvent assez laborieux.

Il ne faut pas perdre de vue cependant que la constante ne donne qu'une valeur par approximation et ne dispense pas du calcul direct si l'on veut arriver à une grande précision. Ainsi, la moyenne de déviation des salaires des ouvriers (adultes mâles) employés dans les ateliers de robinetterie en Belgique (octobre 1903) est 3.234 comme nous l'avons vu au n° 282; la déviation type des salaires des mêmes ouvriers est 4.337 (Cfr. n° 284). Le rapport entre ces deux nombres n'est que $\frac{3,234}{4,337} = 0.745$, inférieur donc à la relation-type donnée plus haut.

La déviation-type (standard-déviation) est la meilleure mesure de la variabilité d'une série. Géométriquement, on peut la définir comme le demi-paramètre de la courbe, c'est-à-dire l'abscisse du point où la courbe de fréquence change de courbure (1).

La précision des résultats se calcule aisément à l'aide de la standard-déviation. On trouve, dit Mansion (2), que la somme des carrés des différences entre un certain nombre de valeurs $x_1, x_2, \ldots x_\mu$ et leur moyenne arithmétique X, est telle que l'on a :

$$\sqrt{\frac{\Sigma\,(X - x)^2}{\mu}} = \sqrt{\frac{\Sigma\,x^2}{\mu} - \left(\frac{\Sigma\,x}{\mu}\right)^2}$$

On peut donc estimer la précision de la moyenne X des μ grandeurs mesurées $x_1, x_2, \ldots x_\mu$ par l'inverse (ou la réciproque) de la racine carrée de la moyenne des carrés des résidus

$$\frac{\Sigma\,(X - x)^2}{\mu}$$

(1) DAVENPORT, *Statistical methods*, p. 16.

(2) BOUDIN-MANSION, *Leçons sur le calcul des probabilités*, p. 279 (3).

La précision de la moyenne est donc donnée (1) par la formule

$$\frac{1}{\sigma} \qquad (31)$$

M. Mansion (2) pense qu'on peut encore mieux exprimer cette précision par la formule

$$\frac{1}{\frac{\Sigma \mid X - x \mid}{\mu}} \qquad (32)$$

des valeurs absolues des différences X — x parce que les valeurs les plus éloignées de la moyenne, qui sont les plus suspectes, y ont moins d'influence. Nous ferons observer cependant que la formule $\frac{1}{\sigma}$ est adoptée généralement en statistique pour exprimer la précision de la moyenne.

IV. — Déviation interquartile.

288. Nous savons qu'on donne le nom de médiane à la valeur centrale de la série statistique. Partant de là, nous sommes amenés à une notion nouvelle : celle d'une valeur qui, de concert avec la médiane, partagerait la série en quatre parties. Admettons que l'on puisse déterminer une valeur telle que tous les termes de la série qui la précèdent ne représentent pas plus du quart du nombre total, et une seconde valeur telle que les observations qui la suivent valent un quart de la somme; alors nous avons le premier et le troisième quartile. On les désigne par les lettres Q_1 et Q_3. Le second quartile se confond évidemment avec la médiane. Les quartiles et la médiane sont des mesures de position qui ont l'avantage d'être simples et faciles à dé-

(1) U. Yule, *loc. cit.*, p. 253.
(2) Mansion, *loc. cit.*, p. 280, note (*).

terminer. L'intervalle compris entre les deux quartiles Q_1 et Q_3 est exactement égal à celui qui précède Q_1 plus celui qui suit Q_3. Dans chacun des quatre groupes constitués par ces mesures, on relève un nombre égal de fréquences et l'on peut poser :

$$\text{Mé} - Q_1 = Q_3 - \text{Mé}$$

dans le cas d'une distribution rigoureusement symétrique. Mais au lieu de cette mesure de dispersion dont les conditions ne se trouvent que rarement réalisées, on admet en pratique la substitution de l'expression suivante :

$$\frac{Q_1 - Q_3}{2} \qquad (33)$$

Pour déterminer Q_1, Q_3 et Mé, on se sert des formules suivantes :

$$\text{Mé} = \frac{n+1}{2} \text{ termes} \qquad (34)$$

$$Q_1 = \frac{n+1}{4} \text{ termes} \qquad (35)$$

$$Q_3 = \frac{3\,(n+1)}{4} \text{ termes} \qquad (36)$$

A l'aide de la même formule, on peut fixer, comme nous le verrons plus loin, la valeur des déciles (dixièmes de la grandeur totale) et les percentiles (centièmes de la série observée).

La médiane et les quartiles sont particulièrement faciles à trouver dans les séries composées de variables classées d'après leur grandeur, chaque grandeur étant répétée autant de fois qu'il y a d'objets la représentant. L'exemple des 94 amandes que nous avons donné en parlant de la médiane comme moyen de représentation synthétique peut servir d'illustration. Le lecteur voudra bien se reporter au tableau dans lequel les 94 amandes sont classées d'après leur longueur (Cfr. n° 259, *ex. 18*). D'après la formule qui

précède, nous savons que la médiane est placée ainsi : $\frac{N+1}{2}$ ou $\frac{94+1}{2} = 47{,}5$, c'est-à-dire qu'elle se trouve entre le 47ᵉ et le 48ᵉ terme; consultant le tableau, nous voyons que les longueurs de la 47ᵉ et de la 48ᵉ amandes sont identiques (3 c. 25) et que la médiane est donc 3 c. 25. Le premier quartile est : $\frac{N+1}{4} = 23{,}7$; en nous reportant au tableau, nous constatons que les longueurs indiquées au nᵒ 23 et au nᵒ 24 diffèrent, étant respectivement 3 c. 00 et 3 c. 02, nous prenons la moyenne de ces deux termes et nous écrivons : $Q_1 = 3{,}01$. Enfin, le troisième quartile est $\frac{3\ (n+1)}{4} = 71{,}1$, c'est-à-dire une valeur intermédiaire entre celles inscrites sous les nᵒˢ 71 et 72; cette valeur étant la même pour les deux variables, nous écrivons : $Q_3 = 3{,}52$.

La mesure de la dispersion interquartile est par conséquent :

$$\frac{Q_1 - Q_3}{2} = \frac{3{,}01 - 3{,}52}{2} = 0{,}255$$

289. La déviation interquartile est particulièrement utile pour caractériser deux séries que l'on désire comparer, par exemple le relevé des salaires à deux époques différentes. Bosco, qui a calculé la déviation interquartile des salaires des ouvriers mineurs belges en 1896 et en 1900 (ouvriers adultes travaillant au fond) d'après l'enquête de l'Office du travail de Belgique (1), a déterminé les résultats suivants (2) :

EXEMPLE 8.

	1896	1900
Quartile inférieur . . . . . . fr.	3.11	4.27
Médiane	3.58	5.07
Quartile supérieur	4.19	6.40

(1) Office du Travail de Belgique, *Statistique des salaires dans les mines de houille*, octobre 1896, mai 1900. Bruxelles, 1901, p. 34.
(2) BOSCO (A.), *Lezioni di statistica*, p. 508.

Vérifions les calculs du statisticien italien en ce qui concerne : 1° la médiane de 1896; 2° le quartile supérieur pour la même année.

En 1896, travaillaient au fond : 61,299 ouvriers. (Cfr. n° 207.)

$$\text{Mé} = \frac{n+1}{2} = \frac{61{,}299+1}{2} = 30{,}650$$

Jusqu'à la classe 3 fr. à fr. 3.49, dont la moitié est fr. 3.25, il y a 155, 861, 2,860, 7,660, 16,456 ouvriers = 27,992. La médiane est donc comprise dans la classe suivante : fr. 3.50 à fr. 3.99. Mais comme cette classe est trop nombreuse (13,444 ouvriers), il y a lieu de procéder à une interpolation simple, du genre de celle que nous avons déjà calculée précédemment. D'après cela, on a :

$$\text{Mé} = 3.49 + \frac{2{,}658}{11{,}444} \cdot {}^1/_2 = 3{,}49 + 0{,}09 = 3{,}58$$

résultat qui concorde avec celui obtenu par Bosco.

$$Q_3 = 3\,\frac{(n+1)}{4} = \frac{183{,}897+1}{4} = 45974{,}5$$

En ajoutant à la classe 3 fr. à fr. 3.49 la classe entière qui la suit immédiatement, on arrive au chiffre de 41,436, proche du troisième quartile. La classe suivante (4 francs à fr. 4.49) est celle qui contient le quartile, mais comme elle est trop nombreuse, il y a lieu de procéder comme ci-dessus. On a donc :

$$3{,}99 + \frac{4538}{11235} \cdot {}^1/_2 = 3{,}99 + 0{,}20 = 4{,}19$$

ainsi que Bosco l'a calculé.

Q_1 calculé d'après la même méthode donne fr. 3.11.

La dispersion interquartile des salaires recueillis en 1896, est de :

$$\frac{3{,}11 - 4{,}19}{2} = 0{,}54$$

La même mesure pour 1900 est :

$$\frac{4,27 - 6,40}{2} = 1,065$$

290. A titre de vérification on peut utilement se rappeler que la mesure de la dispersion interquartile représente environ les deux tiers de la déviation-type, l'exactitude de ce rapport n'existant d'ailleurs que dans les cas où la courbe se rapproche sensiblement de la courbe normale. Ainsi, dans le cas de la distribution des salaires parmi une population ouvrière, on n'obtiendrait pas une vérification complète du rapport dont il s'agit, mais seulement une approximation. La dispersion interquartile des salaires des ouvriers mineurs (fond) en Belgique en 1896 a pour mesure 0.54, pour la même série, σ a pour valeur 0.86.

Le rapport

$$\frac{Q}{\sigma} = \frac{0,54}{0.86} = 0,63 \text{ p. c.}$$

ne répond pas tout à fait à la proportion des deux tiers, parce que la courbe des salaires n'affecte pas la forme de la courbe normale. Malgré ces conditions défectueuses, la règle empirique ci-dessus, en ne la prenant que comme une large approximation, peut être regardée comme un guide de l'esprit et une vérification sommaire du travail arithmétique.

La grande simplicité des opérations arithmétiques nécessaires au calcul assure à la mesure de dispersion interquartile une faveur qu'il serait difficile de lui disputer. Cependant, il ne faut pas s'illusionner sur la valeur du procédé. Il ne prend sa signification entière que si la courbe de la série est à peu près symétrique ou normale. La dispersion interquartile se base sur le fait que, la déviation entre les quartiles étant mesurée, il y lieu de penser que cette valeur représente la série entière; c'est donc comme s'il n'y avait pas de modification sensible à la courbe en

dehors des quartiles. Le postulat est exact pour les courbes normales, mais il est faux pour les autres. Après le troisième quartile, il y a encore place pour de nombreuses altérations. Dans les statistiques de salaires, notamment, on remarque aux époques de hausse une poussée qui étend vers la droite, bien loin au delà du point normal, la courbe de distribution des salaires.

V. — Coefficient de variation.

291. On désigne sous ce nom une mesure relative de la dispersion. Celle-ci est donc comparée à un élément fixe servant de point de comparaison. La quantité obtenue à l'aide de ce calcul est purement abstraite : elle n'a pas de rapport avec l'unité initiale ayant servi à la statistique; ce point est à retenir pour l'interprétation des résultats.

Chacune des mesures que nous avons étudiées sous les numéros précédents a son coefficient de variation relative. On le calcule sur la moyenne arithmétique quand il s'agit de la déviation-type de valeurs non groupées; on se sert de la valeur arbitraire admise pour la moyenne dans le cas de phénomènes se présentant par classes et par fréquences. Le coefficient de variation interquartile se calcule d'après la médiane et il en est de même du coefficient de la moyenne de déviation.

La formule du coefficient de variation a été donnée par M. Pearson sous cette forme en ce qui concerne la déviation-type :

$$V = 100. \frac{\sigma}{M} \qquad (37)$$

dans cette équation, σ est la standard-déviation et M est la moyenne arithmétique.

Pour calculer le coefficient de variation de la moyenne de déviation, on se sert de la formule :

$$V = \frac{\Delta}{Mé} \qquad (38)$$

dans laquelle Δ est le quotient de la division du produit des variables par les écarts et Mé signifie la médiane.

Le coefficient de variation interquartile, s'exprime par la formule ci-après :

$$\frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad (39)$$

dans laquelle Q_3 est le quartile obtenu par la formule

$$\frac{3(n + 1)}{4} \qquad (36)$$

et Q_1 vient de la formule

$$\frac{n + 1}{4} \qquad (35)$$

292. Ces différentes formules étant très simples, il suffira d'en faire de rapides applications :

Moyenne de déviation (coefficient de la)

Exemple 9.

Nombres 19, 25, 26, **28**, 32, 33, 40. Mé = 28. Δ = 5.

$$\frac{\Delta}{\text{Mé}} = \frac{5}{28} = 0{,}1785$$

Exemple 10.

Pimpinella saxifraga L. Mé = 10.305. Δ = 1.429

$$\frac{\Delta}{\text{Mé}} = \frac{1.429}{10.305} = 0{,}13867$$

Dispersion interquartile (coefficient de la)

Exemple 11.

Salaires des ouvriers mineurs belges (adultes travaillant au fond), en 1896 :

$$\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{4{,}19 - 3{,}11}{4{,}19 + 3{,}11} = 0{,}147$$

Déviation-type (coefficient de la)

EXEMPLE 12.

Salaires des ouvriers robinettiers, en 1903 :

$$V = 100. \frac{\sigma}{M} = 100. \frac{4,337}{3,620} = 1,198$$

VI. — Dissymétrie (Skewness).

293. Skewness est le nom, — dont l'équivalent en français pourrait se rendre à peu près par le terme « dissymétrie », — donné par le professeur Pearson à ce fait que la dispersion des variables à partir de la normale n'est pas régulière. Dans la courbe normale on relève que ses points sont symétriques, c'est-à-dire qu'à une égale distance de la moyenne ils sont équivalents. Mais le plus grand nombre de courbes appartiennent à un type asymétrique dérivé de la courbe normale. Il est donc exceptionnel, dans les courbes statistiques, que la moyenne, la médiane et le mode se placent au même point et c'est sur l'éloignement plus ou moins grand de ces mesures l'une par rapport à l'autre que se trouve basée la mesure de dissymétrie ou *Skewness* de la courbe.

Il existe plusieurs procédés pour mesurer la dissymétrie de la courbe.

La mesure la plus simple consiste à faire la différence entre le mode et la moyenne arithmétique.

Soient M la moyenne arithmétique, Mo le mode, la formule :

$$\text{Dissymétrie} = M - Mo \qquad (40)$$

donnera une première mesure de dissymétrie de la courbe. Cette mesure, de même que pour la dispersion, a un coefficient que l'on obtient en divisant la différence entre la

moyenne et le mode par la moyenne de déviation calculée d'après le mode, ainsi que l'indique la formule :

Coefficient de dissymétrie =

$$\frac{M - Mo}{\Delta\ Mo} \qquad (41)$$

mais, à cause de la difficulté de déterminer le mode d'une façon absolument sûre, on substitue souvent la médiane au mode. Les formules de dissymétrie et du coefficient de dissymétrie restent les mêmes, sauf à remplacer la lettre Mo par Mé.

On mesure également la dissymétrie de la distribution des variables autour de la médiane au moyen de l'expression simple :

$$Q_3 + Q_1 - 2\ Mé. \qquad (42)$$

et on la transforme en un coefficient de dissymétrie en la comparant à la déviation interquartile au moyen de la formule :

$$\frac{Q_3 + Q_1 - 2\ Mé}{\frac{Q_3 - Q_1}{2}} \qquad (43)$$

Sans être absolument parfaite, cette expression de l'asymétrie de la distribution possède le mérite d'être claire et facile à calculer. Son exactitude est suffisante pour qu'on l'emploie dans tous les cas où il n'est pas nécessaire de tenir compte des dernières variables.

Mais la mesure la plus généralement adoptée est celle qui rapporte à la déviation-type la différence entre la moyenne et le mode, ainsi que le professeur Pearson a proposé de le faire :

$$\frac{M - Mo}{\sigma} \qquad (44)$$

M. Yule fait observer que la position du mode étant difficile à indiquer avec précision, on peut adopter une formule

approximative basée sur la médiane et la moyenne. Cette formule est la suivante :

$$\frac{3\ (\mathrm{M} - \mathrm{Mé})}{\sigma} \qquad (45)$$

Enfin, on a proposé un coefficient de la dissymétrie basée sur la mise au cube des écarts à la moyenne, au moyen de la formule :

$$\frac{\sqrt[3]{\frac{\Sigma\ d^3}{n}}}{\sigma} \qquad (46)$$

294. Bien que les formules précédentes ne présentent en elles-mêmes aucune difficulté, il ne sera pas inutile de calculer à l'usage du lecteur les résultats de leur application à des données connues.

Les courbes des salaires des ouvriers mineurs en Belgique en 1896 et en 1900 que nous avons utilisées déjà à plusieurs reprises fournissent un matériel excellent pour déterminer la dissymétrie relative de la distribution des variables.

Voyons d'abord les résultats donnés par la formule des quartiles : d'après la statistique des salaires en 1896, $Q_3 =$ fr. 4.19, $Q_1 =$ fr. 3.11, Mé = fr. 3.58. Nous avons donc, en remplaçant la notation de la formule par les données numériques équivalentes,

Exemple 13.

$$\frac{4\ \text{fr.}\ 19 + 3\ \text{fr.}\ 11 - 2 \times (3\ \text{fr.}\ 58)}{\frac{4\ \text{fr.}\ 19 - 3\ \text{fr.}\ 11}{2}} = \frac{0{,}14}{0{,}54} = 0.2592 \qquad (43)$$

Etant donné qu'en 1900, $Q_3 =$ fr. 6.40, $Q_1 =$ fr. 4.27, Mé = fr. 5.07, la même formule donne les résultats suivants :

Exemple 14.

$$\frac{6\ \text{fr.}\ 40 + 4\ \text{fr.}\ 27 - 2 \times (5\ \text{fr.}\ 07)}{\frac{6\ \text{fr.}\ 40 \quad 4\ \text{fr.}\ 27}{2}} = \frac{0{,}53}{1{,}065} = 0{,}4976 \qquad (43)$$

D'après la formule 46 on doit établir la relation entre la différence du mode à la moyenne par rapport à la déviation-type.

Dans les statistiques de salaires, le mode étant souvent difficile à préciser, nous recourrons à la formule approximative donnée par M. Yule pour substituer la médiane au mode. En 1896, la moyenne arithmétique des salaires est fr. 3.67 et $\sigma = 0.86$; en 1900, la moyenne arithmétique est fr. 5.36 et $\sigma = 1.49$.

Ceci étant donné, on a, comme seconde mesure de dissymétrie des courbes :

EXEMPLE 15.

$$(1896) \quad \frac{3 \times (3 \text{ fr. } 67 - 3 \text{ fr. } 58)}{0{,}86} = \frac{0{,}27}{0{,}86} = 0{,}3139 \qquad (45)$$

EXEMPLE 16.

$$(1900) \quad \frac{3 \times (5 \text{ fr. } 36 - 5 \text{ fr. } 07)}{1{,}49} = \frac{0{,}87}{1{,}49} = 0{,}5839 \qquad (45)$$

VII. — Variabilité.

295. Les mesures de dispersion dont nous avons fait l'exposé sont toutes basées sur la moyenne et sur l'écart de chaque terme à la moyenne, cet écart étant considéré comme une erreur dont la mesure la plus exacte est donnée par la standard-déviation (σ).

Ceci nous met sur la voie de considérations desquelles découle l'utilité d'une formule nouvelle, celle de la variabilité.

Les phénomènes considérés sous le rapport de leur dispersion présentent entre eux des différences essentielles.

Comme le dit le professeur Gini (1), il existe des carac-

(1) Prof. Corrado GINI : « *Variabilita e mutabilita.* Bologna, Cuppini, 1912, p. 17.

tères qui, en réalité, conservent, au cours des observations, la même intensité, mais qui se présentent à nous sous différentes modalités quantitatives uniquement à cause des erreurs, accidentelles ou systématiques, commises par l'observateur. En lui-même le fait observé ne change pas, mais la mesure de sa grandeur, de son intensité, n'apparaît pas invariable. Le cas le plus frappant qu'on puisse citer est celui des observations astronomiques : encore qu'elles soient faites par des savants, à l'aide d'instruments d'une grande précision, elles ne sont pas rigoureusement identiques. Il en est des exemples mémorables. Sans doute le phénomène observé reste invariable, mais la faiblesse de nos sens, l'imperfection des instruments que nous employons, certaines circonstances accidentelles ne nous laissent pas arriver à la perfection, au point qu'aucune de nos observations ne s'écartant de la vérité absolue elles seraient toutes rigoureusement semblables. Dès lors, nous n'avons qu'un parti à prendre : choisir, entre ces mesures, celle qui est la plus proche de la vérité, — la moyenne —, puis calculer l'écart de la série par rapport à la moyenne. C'est ce que l'on a fait par le calcul de la moyenne de déviation, de la déviation-type ou standard-déviation, de la déviation interquartile, etc.

296. Mais à côté de ces phénomènes, il en existe d'autres bien plus nombreux qui font l'objet des recherches usuelles de la statistique : ce sont les faits de la vie physique, sociale et économique. Ils ne semblent aucunement se modeler sur un type ; et si nos observations nous les montrent dans un état de perpétuelle évolution, ce n'est pas, comme dans le cas précédent, parce qu'elles sont impuissantes à les saisir avec une précision suffisante. Aussi la question n'est-elle plus : « de combien la quantité x s'éloigne-t-elle de la moyenne ? », mais bien celle-ci, très différente : « de combien diffèrent entre elles les quantités relevées effectivement ? » Les prix variables d'une denrée ne peuvent être

assimilés aux erreurs commises par un astronome dans l'évaluation de l'ascension droite d'un astre. Les prix sont déterminés par un ensemble de facteurs, chaque prix a sa raison d'être, son individualité; nous tenons à savoir si les prix ont beaucoup varié ou non. C'est un problème tout autre que celui qui consiste à savoir quelle est, entre différentes mesures, la plus exacte, la plus probable.

A des problèmes différents, il est rationnel d'appliquer des méthodes différentes. C'est la raison qui a conduit le professeur Corr. Gini (1912) à rechercher une mesure de dispersion autre que celles basées sur la moyenne: cette mesure est le coefficient de variabilité entre n caractères. La question est la suivante : « de combien les diverses grandeurs effectives diffèrent-elles entre elles? » et non « de combien les diverses quantités relevées s'éloignent-elles de leur moyenne arithmétique? » L'indice de variabilité mesure donc l'intensité de la différence entre les grandeurs effectives.

297. Adoptons la notation ci-après: a_1, a_2 ...a_n désignent les n quantités rangées suivant un ordre croissant; a_i est une quantité quelconque de rang i.

Entre deux quantités consécutives de la série, il y a un intervalle d'un grade.

La distance graduelle $d_{i,t}$ qui sépare deux termes a_i et a est égale au nombre de grades qu'il y a entre eux. Les quantités a_i et a_t sont symétriques quand $i + t = n + 1$ ou $t = n - i + 1$ et l'on a $d_{i,n-i+1} = n + 1 - 2i$.

Il y a $n\ (n - 1)$ différences possibles entre les n quantités de la série graduée prises deux à deux et la moyenne arithmétique de ces $n\ (n - 1)$ différences est donnée par la formule :

$$\Delta = \frac{2}{n\,(n-1)} \sum_{i=1}^{\frac{n+1}{2}} (n + 1 - 2i)\,(a_{n-i+1} - a_i) \qquad (47)$$

La somme des $n\ (n - 1)$ différences entre chaque quan-

tité et toutes les *autres* est évidemment égale à la somme des n^2 différences entre chaque quantité et *toutes les quantités,* de telle sorte que la différence moyenne *avec répétition* entre les n quantités sera

$$\Delta_R = \frac{2}{n^2} \sum_{i=1}^{\frac{n+1}{2}} (n+1-2i)(a_{n-i+1} - a_i)$$

et

$$\Delta_R = \frac{n-1}{n} \Delta \qquad (48)$$

Ces deux différences ont chacune leur utilité particulière. Puisque $n + 1 - 2_i = d_{i,n-i+i}$, on peut écrire Δ sous la forme suivante :

$$\Delta = \frac{2}{n(n-1)} \sum_{i=1}^{\frac{n+1}{2}} d_{i,n-i+1} (a_{n-i+1} - a_i)$$

ou encore

$$\Delta = \frac{1}{n(n-1)} \sum_{i=1}^{n} d_{i,n-i+1} \left| a_i - a_{n-i+1} \right|$$

en désignant par $\left| a_i - a_{n-i+1} \right|$ la différence *absolue* entre deux quantités géométriques (1).

298. Le procédé de calcul arithmétique nous intéresse spécialement. On peut le décrire comme suit :

Etant donné un tableau statistique rangeant les variables dans un certain ordre, — que ce soit l'ordre chronologique comme dans certaines statistiques démographiques, ou l'ordre des grandeurs comme dans de nombreuses statistiques sociales et économiques, — on commence par diviser la série en deux parties égales et l'on dresse un tableau dont la première colonne est réservée à la moitié des variables dans l'ordre croissant (de a_1 à $a_{\frac{n}{2}}$) ; les autres sont insérées dans la seconde colonne et classées dans l'ordre décroissant (de a_n à $a_{\frac{n}{2}+1}$). Les termes placés en re-

(1) Cfr. Prof. Corr. Gini, *loc. cit.*, pp. 21-23.

gard l'un de l'autre sont symétriques : on en calcule la différence pour chaque couple $\left| a_{n-i+1} - a_i \right|$ La distance graduelle entre variables symétriques est égale à $n + i - 2i$, soit dans l'exemple proposé $38 + 1 - 2 = 37$, pour $i = 1$; la quatrième colonne est réservée à l'inscription des distances graduelles en ordre décroissant. La cinquième sert à indiquer les produits des différences entre variables symétriques par les distances graduelles $\left[(a_{n-i+1} - a_i) \times (n + 1 - 2i) \right]$. Nous faisons la somme de ces produits et nous multiplions cette somme par $\frac{2}{n\,(n-1)}$ pour obtenir la différence moyenne simple et par $\frac{2}{n\,(n)}$, si l'on veut connaître la différence moyenne avec répétition.

299. Nous faisons application de ces règles de calcul à un cas concret : l'écart entre les importations et les exportations, commerce spécial, en Belgique, de 1876 à 1913. D'abord, nous calculons ces écarts, puis les réduisons à un système sérial d'index-numbers supposant chaque fois les exportations égales à 100 l'année correspondante :

Années	Importations	Années	Importations	Années	Importations	Années	Importations
1876	107.9699	1886	112.9507	1896	121.0353	1906	123.1889
1877	131.8217	1887	115.4203	1897	115.1650	1907	129.6596
1878	132.4008	1888	123.3503	1898	114.4219	1908	130 4250
1879	128.1515	1889	106.7089	1899	115.9519	1909	130.8759
1880	138.1471	1890	116.3595	1090	115 2307	1910	122.2081
1881	125.1177	1891	118.4842	1901	121.4830	1911	122.1327
1882	121.2415	1892	112.1958	1902	123.6403	1912	119.6824
1883	115.5611	1893	116.1653	1903	125.8741	1913	128.1385
1884	106.5993	1894	120.7767	1904	127.4341		
1885	112.2536	1895	121.2905	1905	131.0303		

Les variables comptées au moyen de leurs unités au-dessus de 100 et exprimées pour la simplification du calcul, avec deux décimales, sont rangées dans leur ordre d'importance dans la première et seconde colonne du tableau suivant, dans l'ordre croissant à la colonne 1 et décroissant à la colonne 2.

Exemple 17.

Excédents des importations rangés dans l'ordre		Différences entre les valeurs symétriques	Distances graduelles entre les valeurs symétriques	Produits
croissant de a_1 à $a_{\frac{n}{2}}$	décroissant de a_n à $a_{\frac{n}{2}+1}$	$\lvert a_i - a_{n-i+1} \rvert$	$(n+1-2_i)$	$\lvert a_i - a_{n-i+1} \rvert \times (n+1-2_i)$
6.59	38.15	31.56	37	1167.72
6 71	32.40	25.69	35	899.15
7.99	31.82	25 83	33	786.39
12.19	31.03	18.84	31	584.04
12.25	30.88	18.63	29	540.27
12.95	30.42	17.47	27	471.69
14.42	29.66	15.24	25	381.00
15.17	28.15	12.98	23	298.54
15.23	28.14	12.91	21	271.11
15.42	27.43	12.01	19	228.19
15.56	25.87	10.31	17	175.27
15.95	25.12	9.17	15	137.55
16 17	23.64	7.47	13	97.11
16.36	23.35	6.99	11	76.89
18.48	23.18	4.70	9	42.30
19.68	22.21	2.53	7	17.71
20.78	22.13	1.35	5	6.75
21.03	21.48	0.45	3	1.35
21.24	21.29	0.05	1	0.05

$$\Delta = 6173{,}08 \times \frac{2}{38.37} = 8{,}781052\ldots$$

résultat qui marque la différence moyenne entre quantités sans répétition, tandis que la différence moyenne avec répétition est :

$$\Delta_R = 6173{,}08 \times \frac{2}{38\ 38} = 8{,}549971...$$

L'écart entre les importations et les exportations n'est donc pas constant; il varie, au contraire, dans des limites étendues.

300. Au lieu de la moyenne, nous pouvons aussi utiliser la médiane pour mesurer les écarts de toute quantité à une valeur type. Lorsque le nombre de termes a est un nombre impair, nous avons :

$$\text{Médiane (Mé)} = a_{\frac{n+1}{2}}$$

Lorsque ce nombre est pair, la médiane se place entre

$$a_{\frac{n}{2}} \text{ et } a_{\frac{n}{2}} + 1$$

et peut prendre toutes les valeurs comprises entre ces deux limites.

La différence absolue entre deux quantités symétriques est égale à la somme des différences entre la médiane et chacune de ces quantités :

$$|a_{n-i+1} - a_i| = |a_i - \text{Mé}| + |a_{n-i+1} - \text{Mé}|$$

La distance graduelle entre deux quantités symétriques est égale à deux fois la distance graduelle entre une de ces quantités et la médiane (1)

donc

$$\sum_{i=1}^{n} |a_i - a_{n-i+1}| = 2 \sum_{i=1}^{n} |a_i - \text{Mé}|$$

(1) Cfr. Gini, *loc. cit.*, pp. 27-28.

et comme la distance graduelle entre deux quantités symétriques est égale à deux fois la distance graduelle, on a aussi :

$$\sum_{i=1}^{n} d_{i, n-i} = 2 \sum_{i=1}^{n} d_{i, \text{Mé}}$$

et

$$\sum_{i=1}^{n} d_{i, n-i+1} \left| a_i - a_{n-i+1} \right| = 4 \sum_{i=1}^{n} d_{i, \text{Mé}} \left| a_i - \text{Mé} \right|$$

donc

$$\Delta = \frac{4}{n(n-1)} \sum_{i=1}^{n} d_{i, \text{Mé}} \left| a_i - \text{Mé} \right| \qquad (49)$$

et

$$\Delta_R = \frac{4}{n^2} \sum_{i=1}^{n} d_{i, \text{Mé}} \left| a_1 - \text{Mé} \right| \qquad (50)$$

301. Si plusieurs quantités successives présentent la même valeur, il est préférable de se servir des formules (49) et (50) au lieu des formules (47) et (48).

Soit s les différentes valeurs que prennent les n quantités, x_k ($k = 1, 2, \ldots 5$) une de ces valeurs, f_k le nombre de quantités qui prennent cette valeur, $\left| x_k - \text{Mé} \right|$ l'écart entre x_k et la médiane et $d_{k, \text{Mé}}$ la distance graduelle moyenne de x_k à la médiane (1). Les formules (49) et (50) pourront s'écrire :

$$\Delta = \frac{4}{n(n-1)} \sum_{k=1}^{s} d_{k, \text{Mé}} f_k \left| x_k - \text{Mé} \right| \qquad (51)$$

$$\Delta_R = \frac{4}{n^2} \sum_{k=1}^{s} d_{k, \text{Mé}} f_k \left| x_k - \text{Mé} \right| \qquad (52)$$

moyenne de x à la médiane (1). Les formules (49) et (50) donnent naissance à des calculs un peu plus compliqués

---

(1) La distance graduelle moyenne d'une valeur x_k à la médiane est égale au nombre des cas compris entre x_k et la médiane, $\frac{f_k}{2}$ ou $\frac{f_k + 1}{2}$ suivant que n est pair ou impair, f_k étant le nombre de cas dans lequel se présente la valeur x_k.

que ceux à employer pour les formules (1) et (2). On peut les décrire de la manière suivante:

On dresse un tableau dont la première colonne est réservée à l'indication des grandeurs dans un ordre croissant; en regard, on marque dans une seconde colonne le nombre de fréquences de cette grandeur, et l'on choisit arbitrairement une de ces fréquences comme représentant la médiane Mé. Dans la troisième colonne, on marque l'écart entre une des quantités et la médiane, Mé étant figurée par 0, en augmentant d'une unité à mesure qu'on s'éloigne de la médiane. Le nombre de variables divisé par 2 est l'indice médian; la distance graduelle moyenne d'une valeur x_k à la médiane est égale au nombre des cas compris entre x_k et la médiane, plus $\frac{f_{k+1}}{2}$ ou $\frac{f_k}{2}$ selon que le nombre des n est pair ou impair. Pour connaître la distance graduelle moyenne à porter dans la quatrième colonne, il faut donc faire la différence de chaque quantité à la valeur médiane, en calculant le nombre de cas compris entre cette quantité et la médiane, et y ajouter le nombre de variables plus 1, divisé par 2 si le nombre des n est pair. Enfin on effectue les produits de la distance graduelle moyenne à la médiane par l'écart et par le nombre de fréquences.

302. Faisons application de la formule ci-dessus à l'exemple 28 du chapitre II (Cfr. n° 185), dans lequel 94 amandes sont mesurées dans le sens de la longueur, à un dixième de millimètre près. Afin d'obtenir une répartition par classes, nous avons divisé les variables en 16 catégories correspondant à des mesures variant d'un millimètre (*voir tableau ci-après*).

EXEMPLE 18.

Dimensions en millimètres x_k (1)	Nombre d'amandes f_k (2)	Écart à Mé $\lvert x_k - \text{Mé} \rvert$ (3)	Distance graduelle moyenne (*) $d_{k,\text{Mé}}$ (4)	PRODUITS $\lvert d_{k,\text{Mé}} f_k \rvert x_k - \text{Mé} \rvert$ (5)
25	2	7	46	644
26	5	6	42.5	1275
27	2	5	39	390
28	2	4	37	296
29	11	3	30.5	1006.5
30	8	2	21	336
31	12	1	11	132
32	14	0	—	—
33	10	1	14	140
34	4	2	21	168
35	6	3	26	468
36	9	4	33.5	1206
37	4	5	40	800
38	2	6	43	516
39	2	7	45	630
40	1	8	47	376

$$\sum_{k=1}^{x} \left| d_{k,\,\text{Mé}} f_k \right| x_k - \text{Mé} \left| = 8383{,}5 \right. \qquad (51)$$

$$8383{,}5 \times \frac{4}{94.93} = 3{,}83596$$

C'est la différence moyenne entre quantités, sans répétition; multiplié par

$$\frac{4}{n^2}, \left(8383{,}5 \times \frac{4}{94.94}\right)$$

c'est la différence moyenne avec répétition.

* Exemple. — Nombre $n = 94$; Mé $= \frac{94}{2} = 47$. Exemple du calcul : classe 31; variables ayant cette dimension et moins $= 42$; donc $47 - 42 + \frac{12}{2} = 11$; $11 \times 1 \times 12 = 132$.

303. La comparaison des degrés de variabilité de certains phénomènes conduit parfois à des conclusions fort intéressantes. Il en est ainsi, par exemple, dans les recherches anthropométriques. Le degré de variabilité de l'indice céphalique, plus ou moins élevé dans certaines parties d'un même pays, montre clairement l'existence de plusieurs races sur le sol patrial. D'après les résultats consignés par M. Livi dans son « Anthropométrie militaire » (1898), M. le docteur G. Dettori a calculé les différences moyennes entre indices céphaliques des hommes astreints au service militaire, répartis d'après leur province d'origine. Voici le résultat de ces recherches (1) :

Provinces	Différence moyenne entre indices	Numéros d'ordre	Provinces	Différence moyenne entre indices	Numéros d'ordre
Vénétie	4.17	14	Latium	4.73	4
Lombardie	3.95	16	Abruzzes	4.61	5
Piémont	4.39	8	Pouilles	4.48	6
Ligurie	4.26	13	Campanie	4 26	12
Emilie	4.38	9	Sardaigne	4.37	10
Toscane	5.06	2	Basilicate	4.47	7
Marche	5.19	1	Calabre	4.33	11
Ombrie	4.74	3	Sicile	4.15	15

D'après les chiffres ci-dessus, on remarquera une variabilité plus grande dans les provinces centrales que dans les provinces du nord et du sud. Cette constatation, fait observer M. Gini, est d'accord avec les théories anthropologiques d'après lesquelles les populations du nord et du sud de l'Italie appartiennent à deux races distinctes qui se seraient fusionnées ou juxtaposées dans la partie centrale de la péninsule.

(1) D'après le tableau reproduit dans l'ouvrage cité du professeur Corrado Gini, p. 31.

304. — *Références.*

Bowley (A. L.), *Elements of statistics*, 2e édit. London, 1902, p. 136.

Elderton (W. Palin), *Frequency-curves and correlation.* Londres, Layton, 1906.

Galton (F.), *Statistics by Intercomparison*. Phil. Mag., vol. XLIX, quatrième série, 1875, pp. 33-46.

Id., *Natural inheritance*. London, 1889, passim.

Gini (Prof. Corrado), *Variabilita e mutabilita; contributo allo studio delle distribuzioni e delle relazioni statistiche.* Bologna, Cuppini, 1912. (Estratto dagli Studi economico-giuridici della R. Universita di Cagliari, anno III, Parte 2a.)

King (W. J.), *The elements of statistical methods.* New York, 1912, ch. XII, XIII et XIV.

Mortara (G.), « Metodi elementari per lo studio delle distribuzioni di caratteri » (*Giornale degli Economisti e rivista di statistica*, 2e sem. 1910, p. 9).

Pearson (K.), *Contributions to the mathematical theory of evolution.* Phil. Trans. Roy. Soc., série A, vol. CLXXXV, 1894, p. 80.

Id., *Skew variation in Homogeneous material*. Phil. Trans. Roy. Soc., série A, vol. CLXXXVI, 1895, p. 343.

Id., *Regression, Heredity and Panmixia.* Phil. Trans. Roy. Soc., série A, vol. CLXXXVII, 1896, p. 253.

Yule (G.-N.), *An introduction to the theory of statistics.* London, p. 132 et suivantes.

Id., « Notes on the History of Pauperism », *Journ. of the Statist. Soc.*, 1896, p. 330.

Zizek, *Statistical averages*, trad. anglaise de Warren Milton Persons. New York, Holt and Co., 1913, part. II, ch. II, p. 256.

CHAPITRE IV

Covariation (*Corrélation*).

I. — Portée du coefficient de covariation ; sphère d'application ; examen critique général.

305. Depuis un certain nombre d'années on a fait usage, spécialement dans les sciences biologiques, en anthropologie et en botanique, de formules mathématiques ayant pour objet d'apprécier l'étroitesse des rapports — ressemblance ou dissemblance — existant entre deux faits distincts. Ces formules ont pour but de représenter schématiquement une quantité de constatations qui pourraient être faites à propos de chacune des données numériques des séries. Il s'agit bien ici, encore une fois, de ces procédés d'analyse qui, appliqués aux résultats des investigations scientifiques, résument, en une expression abrégée, la multitude des chiffres et servent de guide et de soutien à l'esprit. Cette réflexion n'est pas inutile au seuil de la matière difficile qui nous occupe, car elle nous rappelle que les symboles ne peuvent avoir une signification plus étendue ou plus profonde que les phénomènes qu'ils représentent : exprimés au moyen des formules d'où se tire le coefficient r, les faits que l'on compare ne sont pas plus étroitement liés l'un à l'autre que quand on les considère dans leurs données simples et qu'on recherche, par les procédés logiques, les rapports qui peuvent exister entre eux. Ce n'est pas parce que le coefficient sera élevé que l'on pourra, d'une manière absolue, conclure à l'étroitesse des liens existant entre les deux séries mises en équation. Le coefficient r a donc une fonction représentative ; il importe, avant tout,

pour l'homme de science, de ne pas se laisser entraîner loin du terrain solide des faits et de ne pas oublier que les symboles statistiques, création de notre esprit, ne peuvent valoir contre les réalités.

A une époque plus récente, les procédés de calcul dont nous venons de parler ont été étendus à la sphère des phénomènes économiques et moraux, quoique avec assez d'hésitation ; les applications de ces méthodes mathématiques, dans ce domaine, étaient, en 1909, encore assez rares pour que M. Udny Yule eût la possibilité de les énumérer toutes dans un mémoire succinct adressé à l'Institut international de statistique (1). Bien que ce fait ne prouve rien contre la valeur des méthodes — valeur du reste incontestable, sauf les abus qui peuvent être commis parfois — il est assez significatif d'autre part quant à la difficulté de trouver un terrain qui se prête sans contestation à l'application du coefficient.

Les formules dites de « corrélation » présentent une connexion étroite avec la théorie des erreurs, spécialement avec la méthode des moindres carrés (2). C'est pourquoi on peut, à ce point de vue spécial, faire remonter à Bravais (1846), les origines de la méthode, car la formule de la somme des produits peut être regardée en un certain sens comme étant due à ce savant. Mais Sir Francis Galton fut le premier à imaginer la méthode statistique pratique (1886-1888) et c'est à lui que nous devons l'idée d'une mesure numérique de la grandeur de corrélation. Edgeworth et Pearson ont amplifié et perfectionné les procédés

(1) L'exposé des applications nouvelles de la méthode des covariations mériterait une étude spéciale, qui ne serait pas dépourvue d'intérêt. Nous n'avons pu l'entreprendre parce qu'un travail de ce genre ne rentre pas dans le cadre d'un ouvrage général tel que le nôtre, mais nous souhaitons que ce sujet tente, sans tarder, l'un ou l'autre chercheur.

(2) Cfr. Udny Yule : « Les applications de la méthode de corrélation aux statistiques sociales et économiques ». (*Bulletin de l'Institut International de statistique*, t. XVIII, 1909, traduction française, pp. 265-277, texte anglais, p. 537.)

de Galton; c'est Pearson qui a introduit dans la pratique (1896) la formule de la somme de produits pour exprimer le coefficient r : enfin, de nombreux savants, dans une série de travaux remarquables, ont fait de nombreuses applications de la méthode, l'ont encore développée et perfectionnée et ont contribué pour une large part à la faire connaître dans le monde des économistes et des statisticiens (1).

306. En règle très générale, la méthode est connue sous le nom de méthode de corrélation.

Avec M. Lucien March, nous regrettons que ce terme soit entré dans le vocabulaire statistique, à cause du sens qu'il a en logique et dans le langage courant, sens qui n'est pas celui qu'il a en statistique.

Lorsqu'on dit qu'un fait est en corrélation avec un autre, on veut exprimer habituellement qu'un de ces faits est dans une relation telle avec l'autre, que l'un suppose l'autre (Littré). En logique, le caractère corrélatif de deux phénomènes implique l'existence entre eux d'un lien causal.

Certains auteurs, qui se sont attachés au côté mathématique de la question plutôt qu'à son aspect philosophique, n'ont pas hésité à écrire que l'existence d'une corrélation, d'un degré assez élevé, entre deux phénomènes supposait l'existence du principe de causalité (2). D'aucuns se sont montrés moins affirmatifs et se sont bornés à dire que des fluctuations dans le même sens ou dans un sens opposé, en grand nombre, démontraient qu'entre les deux phénomènes il existait une relation dont le caractère plus ou moins étroit est mesuré par le coefficient de corrélation. Cette opinion, bien que plus modérée que la première, n'est pas admise par M. March qui écrit ceci : « le résultat d'un cal-

(1) Citons parmi eux les statisticiens anglais, italiens et américains, spécialement MM. L. Bowley, G. Udny Yule, G. Mortara, C. Gini, Turroni Bresciani, etc.

(2) King, *Elements of statistical methods*, n° 109, p. 197.

cul ne peut mesurer la relation effective de deux séries de phénomènes ; ce résultat peut accuser, en effet, une relation apparente très étroite alors que la relation réelle est nulle ou de sens opposé à la relation apparente. La relation de deux séries de grandeurs peut d'ailleurs s'entendre de diverses façons : si l'on décompose, par exemple, le mouvement des termes d'une série chronologique en changements à longue et changements à courte période, la relation peut avoir un certain sens pour les premiers, un autre sens pour les seconds (1). »

Nous partageons l'avis de M. L. March et nous pensons avec lui que l'on a dépassé la portée des conclusions légitimes lorsqu'on a donné à la méthode le nom de corrélation. Celui de covariation, qui implique seulement l'existence d'une concomitance entre les variations des phénomènes, est proposé et employé par M. March, et nous adoptons cette terminologie pour les motifs indiqués plus haut.

307. Si le coefficient de covariation ne peut nous renseigner exactement au sujet de l'existence d'un lien causal entre deux phénomènes que l'on compare, quelle est donc au juste sa fonction et sa signification? Est-ce un jeu de l'esprit, une superfétation mathématique, une satisfaction donnée à une mode momentanée ou au désir de paraître plus savant, d'imprimer à l'œuvre statistique un cachet plus rigoureusement scientifique qu'on ne le pourrait à l'aide des raisonnements de la simple logique? Le prétendre serait singulièrement et fort injustement rabaisser la portée théorique et pratique de la méthode. En effet, les calculs par lesquels se détermine le coefficient de covariation ont, sans conteste, une grande utilité. En premier lieu,

---

(1) MARCH (L.), « De l'application de procédés mathématiques à la comparaison des statistiques ». (*Bull. Instit. Intern. Stat.*, vol. XVIII, 1909, p. 262.)

l'indice de covariation précise le degré de parallélisme ou d'antiparallélisme des courbes. Il ne suffit pas de constater que les séries sont formées de nombres dont la succession est à peu près la même de part et d'autre, ou est en sens opposé; il faut encore préciser ces ressemblances ou ces divergences. Quiconque a essayé d'interpréter un matériel statistique s'est heurté, l'une ou l'autre fois, à une difficulté de l'espèce. Or, le coefficient de covariation donne l'expression la plus exacte possible en cette matière. Ce premier avantage est déjà notable.

Il nous semble que, sous un autre rapport, les calculs de l'espèce ont une utilité qu'on ne peut non plus sous-évaluer : celle d'orienter les réflexions du chercheur dans un sens déterminé. Au début du processus du raisonnement logique, l'esprit hésite entre plusieurs hypothèses plausibles : existerait-il un rapport entre tel phénomène et tel autre? Au contraire, la constatation du fait A signifie-t-elle que le fait B ne se produira pas? Le coefficient de covariation peut fournir une première réponse à cette question. Une valeur élevée de ce coefficient doit nous faire réfléchir à la possibilité d'une union entre phénomènes, même éloignés l'un de l'autre; dans le sens opposé, une valeur faible du même coefficient pourra nous dissuader d'associer trop intimement deux ordres d'idées que nous serions enclins à trouver voisins.

Enfin, le coefficient de covariation a un rôle à jouer dans la vérification de nos raisonnements et ce n'est pas là un office dont l'importance soit à dédaigner.

Ces points de vue sont excellemment résumés par un auteur déjà cité lorsqu'il écrit : « Dans l'enchaînement des opérations logiques de l'esprit, l'intervention des procédés mathématiques dans les comparaisons statistiques a une utilité évidente aux extrémités de la chaîne : au point de départ pour suggérer des conclusions possibles, ou à la fin pour contrôler la partie analytique du processus : elle ne

peut, en aucune façon, constituer le nœud même du raisonnement (1). »

Ces derniers mots sont à retenir. Les formules mathématiques — et celles de la covariation en premier lieu — ne dispensent pas de raisonner; lorsque la valeur de r est déterminée, on n'a pas tout fait. Quelques-uns ont paru penser que le raisonnement logique pouvait être remplacé par l'emploi des formules appropriées à la recherche des connexions entre les séries statistiques; dans notre opinion, cette prétention est injustifiable et aboutirait à des confusions regrettables.

Le coefficient de covariation est essentiellement un guide pour l'esprit, au début et à la conclusion de la recherche scientifique. On peut même aller plus loin et dire, avec M. Udny Yule que la grandeur du coefficient de covariation montre l'amplitude de l'écart à attendre dans un phénomène donné, par rapport à la moyenne, pour un certain écart constaté dans un phénomène connexe, à condition qu'on puisse juger de l'avenir d'après les données du passé (2) : cette conclusion peut être étendue à tous les coefficients, même calculés par les voies les plus simples. Quant à la connexion causale, on peut dire que le parallélisme des courbes démontre la concomitance des mouvements représentés et qu'il permet seulement de soupçonner la connexité, la dépendance des circonstances qui déterminent les mouvements (3).

308. La sphère d'application de la méthode de covariation aux phénomènes sociaux et économiques est, avons-nous dit, assez étroite. Nous ne reprendrons pas ici l'exa-

---

(1) March (L.), « De l'application des procédés mathématiques à la comparaison des statistiques ». (*Bull. Instit. Intern. Stat.*, vol. XVIII, 1909, p. 261.)

(2) Yule (G. U.), « Les applications de la méthode de corrélation aux statistiques sociales et économiques ». (*Bull. Instit. Intern. Stat.*, t. XVIII, 1909, trad. franç., p. 266.)

(3) March (L.), « Comparaison numérique des courbes statistiques ». (*Journal de la Société de Statistique de Paris*, 1905, p. 256.)

men détaillé auquel il a été procédé par M. Udny Yule dans son mémoire à l'Institut international de statistique, exposé qui devrait être complété. Cependant le lecteur doit connaître l'étendue du domaine auquel se sont attachées, en matière économique et sociale, les premières recherches du coefficient de variation et force nous est de répéter, en partie et sous une forme succincte, certaines choses que l'on a dites avant nous.

Les premiers exemples d'application de la formule de covariation à des phénomènes économiques remontent à 1895 et 1896; ces recherches, faites par M. Udny Yule, avaient pour objet de vérifier si l'assistance des pauvres à domicile, dans l'organisation de la bienfaisance publique en Angleterre, avait une influence marquée sur l'accroissement du paupérisme. Des recherches sur le même objet, plus développées, et dans lesquelles les variations du paupérisme étaient mises en rapport avec d'autres éléments, furent publiées par le même auteur quelques années plus tard.

Une question ayant fait l'objet de fréquentes recherches est celle des rapports qui existent entre la nuptialité et les circonstances économiques, telles que le développement du commerce, des affaires ou le bon marché de la vie (mesuré approximativement par les variations du prix du blé). De nombreux économistes ont consacré des études à ce problème et tous, à peu près, ont conclu à une action des conjonctures économiques sur le taux de la nuptialité. M. Hooker a entrepris de donner une forme mathématique, celle du coefficient de covariation, à des recherches analogues. Son travail, qui a introduit une méthode nouvelle de mesurer les écarts en les ramenant à la moyenne d'une courte période et non à la moyenne de la période entière, conclut à l'existence d'une liaison entre le taux de la nuptialité et le mouvement dans les « clearing-house » quelque quinze mois auparavant, tandis que la connexité avec le prix du blé paraît très faible.

A son tour, M. Lucien March, s'aidant d'une méthode simple que nous analyserons plus loin, a fait des applications de la méthode de covariation, comme il l'a appelée lui-même, à des questions démographiques et économiques, telles que les fluctuations du taux des mariages et des naissances et du coefficient de nuptialité et du chômage. M. Yule, un peu plus tard, sans connaître le travail de M. March, a traité des données analogues et a longuement discuté la nature réelle de la relation entre les fluctuations du taux du mariage et du commerce.

La fécondité et le milieu social ont déjà été analysés par différents économistes et démographes. M. Héron a repris l'étude du problème sous son aspect mathématique.

On a encore essayé de mesurer l'influence du temps sur les récoltes et les conclusions de M. Hooker sur ce sujet ne manqueront pas d'intéresser les agronomes et les économistes.

Les questions boursières ont été analysées par M. Hooker dans ses deux études successives sur l'effet de la suspension de la Bourse de Berlin et par M. Norton (rapport des réserves et de l'escompte).

Enfin, on s'est demandé si le taux d'accroissement de l'importation en Angleterre, des objets manufacturés avait une influence sur le nombre de membres non chômeurs dans certaines « Trade-Unions ». Nous aurons plus loin l'occasion d'analyser, au point de vue critique, cette application de la méthode de covariation. (Cfr. n° 30.)

Depuis que le mémoire de M. Yule a paru, de nouvelles applications de la méthode ont été faites à d'autres sujets du même genre (1). Les indications qui précèdent n'ont pas

---

(1) Signalons, par exemple, l'étude de M. Marcel Lenoir : « Prix, production et consommation de quelques marchandises (charbon, blé, coton, café) », parue dans le *Bulletin de la Statistique générale de la France*, t. I (octobre 1912-juillet 1913), Paris, Alcan, pp. 172-214. Dans cet important mémoire l'auteur a calculé le coefficient de corrélation entre le stock existant en charbon sur le carreau de la mine et le prix de la tonne (= — 0.71), entre les oscillations

le caractère d'une notice bibliographique; elle n'ont d'autre objet que de faire connaître au lecteur les sujets qui ont paru s'adapter le mieux au calcul de la covariation. On aura déjà remarqué que les principales relations entre deux ou plusieurs phénomènes économiques avaient été indiquées longtemps avant qu'on eût songé au coefficient de covariation. William Farr, par exemple, bien longtemps auparavant, considérait les oscillations du taux des mariages comme une sorte de baromètre de la situation économique. La relation entre l'encaisse et l'escompte était classique longtemps avant qu'on n'eût prononcé le mot de coefficient de covariation. Ce n'est donc pas une invention que nous devons à l'emploi de cette méthode mathématique, mais une confirmation et une précision. Le coefficient de covariation joue exactement son rôle de guide de l'esprit : il confirme et précise les conclusions simples basées sur la logique, sans remplacer l'effort du raisonnement.

309. Comme le fait très justement observer M. Yule, le coefficient de covariation exprime, en une forme sommaire et compréhensible, un aspect particulier des faits sur lesquels il est basé, de sorte que les difficultés véritables commencent quand il s'agit d'interpréter le coefficient ainsi obtenu (1). En d'autres termes, tout dépend de la correction avec laquelle les faits qu'on désire comparer sont associés. Si cette association a été faite suivant les principes logiques, on obtient un coefficient présentant une valeur

---

du prix et de la consommation (= + 0.59 entre 1876 et 1905), entre les quatre variables suivantes : prix moyen de la houille en Europe, accroissement moyen de la production de la houille en Europe, production moyenne de métal monnaie, et le temps (prix et accroissement de la production = — 0.34; — prix et production du métal monnaie or et argent = + 0.73, or seul = + 0.82. Des calculs analogues sont établis pour les autres marchandises. On trouvera p. 214 de la publication susmentionnée un Appendice renfermant un exemple du calcul des coefficients de corrélation et des équations de régression.

(1) Yule (U. G.), *An introduction to the theory of statistics*, p. 191.

sérieuse ; dans le cas contraire, on aboutit au résultat opposé. C'est pourquoi, avant de procéder aux calculs assez laborieux qu'exige la détermination du coefficient de covariation, il importe de bien préciser la nature des données et de vérifier si la comparaison de certains éléments n'est pas plus suggestive et plus exacte que celle qui porte sur d'autres points. La nécessité de cet examen prouve, une fois de plus, que le nœud du raisonnement ne consiste pas dans l'application des formules mathématiques.

Essayons de marquer cette nécessité en discutant les éléments à rapprocher dans l'étude d'une des questions signalées dans la revue sommaire que nous avons faite des applications de la méthode de covariation aux problèmes économiques et sociaux. (Cfr. n° 308.) Il s'agit de savoir si l'augmentation de la valeur, d'une année à l'autre, des articles entièrement ou en majeure partie manufacturés, importés en Angleterre, a une influence sur la proportion des membres non-chômeurs des «Trade-Unions». L'auteur (1) pose ainsi le problème : « que faut-il penser de l'argument des partisans de la réforme du tarif, consistant à dire : l'importation de produits manufacturés favorise l'emploi de la main-d'œuvre à l'étranger, au lieu des ouvriers anglais ? » Se basant sur un rapport du « Board of Trade » (1906) relatif au marché du travail et à l'importation de produits manufacturés, l'auteur, dans un tableau portant sur les années 1860 à 1904, établit : *a*) le chiffre de l'importation (valeur) des produits manufacturés (en milliers de livres) ; A') le taux d'augmentation de la valeur des articles importés par rapport à l'année précédente ; *b*) le pourcentage des ouvriers au travail, d'après les renseignements fournis par les « Trade-Unions » ; B') le taux des changements survenus

(1) Lee (Miss Alice) « On the manner in which the percentage of employed workmen in this country is related to the import of articles wholly or mainly manufactured ». (*Economic Journal*, 1908, vol. 18, p. 96.)

dans le pourcentage des ouvriers au travail (1). L'auteur a calculé le coefficient de corrélation entre A' et *b* et entre A' et B'; il obtient respectivement 0.31 ± 0.09 et 0.47 ± 0.08. La conclusion est qu'une importation plus considérable d'articles manufacturés est concomitante avec une proportion plus considérable d'hommes au travail et qu'une diminution de ces importations augmente le manque de travail (2).

310. Le procédé de l'auteur étant ainsi décrit, nous nous poserons la question de savoir si le problème était susceptible d'une solution, étant donnés les éléments dont on disposait, et si les éléments à soumettre au calcul étaient bien choisis.

En ce qui concerne le premier point, des doutes sérieux peuvent être élevés à raison des considérations ci-après :

A. — Nature des produits compris sous la dénomination générale : *Objets manufacturés.* Dans les statistiques commerciales, indépendamment des dénominations s'appliquant aux produits ou aux classes de produits, on trouve généralement des nomenclatures plus larges qui ont pour objet de donner une idée de la nature, dans leur ensemble, des transactions. En Angleterre (3), les quatre groupes suivants existaient à l'époque envisagée : 1° substances alimentaires, boissons et tabac; 2° matières premières et articles complètement bruts; 3° articles complètement ou partiellement manufacturés; 4° articles divers et non classés. Les

(1) Obtenu en divisant la différence entre la première année et la seconde par la proportion du nombre d'ouvriers au travail la seconde année. Ex. : année 1860 = 98.15; 1861 = 96.30; taux cherché $= \dfrac{98.15 - 96.30}{96.30} = \dfrac{1.85}{96.30} = 0.0192$. (Cfr. Miss Alice Lee, *loc. cit.*, pp. 99).

(2) Lee (Miss Alice), *op. cit.*, *loc. cit.*, p. 99.

(3) Voyez un résumé de la statistique commerciale anglaise dans notre *Précis du cours de statistique générale et appliquée*, 3e édit., Bruxelles, 1912, nos 206 et 207.

données utilisées *sub litt.* A visent les articles compris dans la troisième division.

Il est de toute évidence que sous cette rubrique, tout à fait générale, sont comptées beaucoup de marchandises disparates. Combien y en a-t-il, parmi elles, qui intéressent le travail d'ouvriers anglais? On ne le sait, mais ce qu'on peut affirmer c'est que maints produits de luxe ou de fantaisie, qui ne sont pas fabriqués en Angleterre, trouvent dans ce pays un marché très avantageux. Il n'est pas possible de prétendre que l'importation de ces objets nuise, en aucune façon, à l'ouvrier anglais en le privant de son travail, puisqu'il s'agit d'industries qui n'ont pas ou presque pas d'équivalent en Grande-Bretagne. Quant à soutenir que l'argent dépensé pour ces importations aurait appauvri le fonds des salaires et de la sorte, aurait empêché de donner de l'ouvrage à un plus grand nombre de travailleurs, on sait que cette théorie du fonds des salaires, tout à fait démodée aujourd'hui, n'a plus besoin d'être réfutée.

Le rapport entre l'état du marché du travail et l'importation d'objets manufacturés ne peut être recherché avec chance de succès que si l'on compare des industries homogènes, non si l'on s'en tient à des généralités. Les termes trop généraux dans lesquels le problème est posé ne permettent pas de lui donner une solution certaine par l'emploi de la méthode de covariation.

B. — Les changements de valeur, d'une année à l'autre, des importations de produits manufacturés, ont été choisis par l'auteur pour être mis en corrélation avec le pourcentage d'ouvriers au travail. Cette méthode n'échappe pas à l'inconvénient de faire dépendre le taux d'augmentation des mouvements annuels et séculaires des prix. Pendant une période de dépression économique l'augmentation de valeur sera arrêtée ou même fera place à une réduction; au contraire, au cours d'une période d'expansion, on verra les accroissements annuels atteindre un taux élevé. Les deux

cas peuvent se produire sans qu'il y ait aucun changement effectif dans les quantités.

Il nous semblerait préférable de calculer la part relative des importations d'objets manufacturés dans le total des importations d'une année et de comparer, d'une année à l'autre, les modifications survenues dans ces proportions. Cette méthode échappe presque entièrement aux influences des changements de prix parce que l'on base le calcul des parts relatives sur les résultats d'une seule année.

Il eût été désirable de n'utiliser que les données exprimant le pourcentage d'ouvriers au travail dans des métiers produisant des objets manufacturés, au lieu de choisir ceux qui se rapportent à la masse ouvrière tout entière. La précaution était d'autant plus recommandable que, pays de grande industrie. l'Angleterre compte un grand nombre d'ouvriers occupés à la préparation de matières premières ou d'articles complètement bruts.

C. — Il est enfin à remarquer que la variation du mouvement du commerce est de courte période; comme le fait remarquer M. Yule, dans une période de commerce très intense, ses importations de tout genre s'élèvent et le travail s'accroît. Nous pensons avec notre éminent collègue que « la méthode ne paraît pas propre à élucider l'existence ou la non-existence d'une relation séculaire ».

II. — Indice de dépendance.

311. Il n'en reste pas moins vrai que l'esprit tient à posséder des notions précises sur le degré de concordance existant entre deux séries ou deux courbes à mettre en comparaison. Les procédés qui visent à atteindre ce résultat ont une utilité analogue à celle de la moyenne. Cette mesure est une expression synthétique d'une série; les indices de covariation jouent le même rôle quant à la ressemblance de deux ou de plusieurs séries. Il ne faut donc pas s'étonner

que des essais nombreux aient été faits en vue de trouver une expression synthétique suffisamment compréhensive, sensible et exacte. Parmi les auteurs qui ont abordé ce problème, il faut citer Fechner qui est l'auteur de l'indice du coefficient de dépendance et ensuite M. March (1) qui a exposé la matière avec une grande clarté et dont nous utiliserons plus d'une fois le savant travail.

Cette détermination est basée sur l'idée très simple de la comparaison des courbes statistiques.

Lorsque deux faits statistiques sont exprimés au moyen de courbes, on se trouve porté, comme malgré soi, à comparer ces courbes entre elles et à évaluer les différences et les ressemblances qu'elles présentent. Si les deux courbes ont une allure identique, si l'une s'élève quand l'autre s'élève, et descend quand l'autre descend, on dira qu'elles sont parallèles et que les phénomènes qu'elles expriment sont liés entre eux, ou sont dominés par une cause extérieure qui les conditionne tous deux. Dans l'ordre des choses économiques et sociales, ce parallélisme parfait est d'une telle rareté qu'il est permis de dire que, pratiquement, il n'existe pas. Mais, sans atteindre ce degré d'union avec un phénomène, beaucoup de séries statistiques sont liées entre elles de telle façon que la connexité des faits ne peut être mise en doute; sans que le parallélisme soit parfait, la tendance générale est la même et parfois, pendant plusieurs années, les courbes se rapprochent du parallélisme complet. Dans un sens opposé, il peut y avoir entre deux séries opposition absolue, c'est-à-dire qu'à tout mouvement ascensionnel d'une courbe correspond une baisse de l'autre et inversement. Cet antiparallélisme trahit encore la dépendance des deux faits; ce rapport, si les deux courbes étaient rigoureusement antiparallèles, serait aussi étroit que si elles étaient absolument parallèles, car il suffirait de re-

---

(1) March (L.), « Comparaison numérique des courbes statistiques ». (*Journal de la Société de Statistique de Paris*, 1905, p. 255.)

tourner l'une des courbes sur elle-même pour qu'elle s'applique exactement sur la seconde. Ce cas d'opposition complète est tout aussi rare que le précédent. La dépendance absolue des grandeurs comparées est dite positive quand les courbes sont parallèles, elle est dite négative dans le cas contraire; elle a l'unité pour expression mathématique.

Les cas d'accord ou de désaccord absolu appartiennent au domaine de la théorie plus qu'à celui de la pratique. Dans les sciences d'observation et surtout dans les sciences sociales, ils sont à peu près inconnus. Le plus souvent on se trouve devant des séries qui accusent une dépendance partielle. Il n'est pas possible d'arriver à quelque exactitude par la simple inspection des courbes. L'observateur le plus attentif doit renoncer à exprimer, à la simple vue d'un diagramme, le degré de dépendance des phénomènes dont l'évolution dans le temps est retracée par le graphique. A plus forte raison est-il impuissant à établir des comparaisons entre deux graphiques différents. Aussi a-t-il fallu trouver un moyen simple de traduire en une donnée numérique la liaison des faits considérés. Un premier moyen est la recherche de ce que M. March, dans son travail cité plus haut, a appelé l'indice de dépendance.

312. La méthode consiste à comparer les données relatives à chaque année aux données concernant l'année suivante en affectant du signe + toute variation positive et du signe — toute variation négative. Si aucune variation ne se marque d'une année à l'autre, on inscrit le chiffre 0. Cette première partie du travail étant achevée, on fera le produit des signes relatifs à chacune des deux séries; ainsi, toute double variation de même signe, positif ou négatif, sera inscrite comme variation positive, et toute double variation de signe contraire sera considérée comme une variation négative. Partant de ces données, on évaluera l'indice en faisant la différence des signes positifs et négatifs et en

divisant cette différence par la somme des intervalles, non compris ceux où la variation est nulle. La formule de l'indice de dépendance est donc :

$$\frac{c - d}{n} \qquad (53)$$

dans laquelle c marque le nombre de variations positives, d celui de variations négatives, n le nombre d'intervalles, On peut aussi écrire :

$$\frac{c - d}{c + d} \qquad (54)$$

lorsque les deux séries ne présentent aucun produit nul.

Comme le fait observer M. March (1), une discordance détruit une concordance. Dans toute concordance, les deux variations sont de même signe, et le produit est positif; dans toute discordance, les variations sont de signe opposé et le produit est négatif. En formant la somme algébrique, les produits négatifs se retranchant des produits positifs, les discordances annulent en quelque sorte automatiquement des concordances en nombre égal. Le dénominateur est une expression positive que l'on réduit, pour simplifier, au nombre des intervalles.

M. March a fait une application très intéressante de cette méthode à la comparaison de la nuptialité à la natalité et au sens des changements annuels de divers comptes du bilan de la Banque de France. Nous ferons nous-même d'autres applications à des phénomènes économiques observés en Belgique.

313. Dans un travail antérieur auquel nous avons déjà eu l'occasion de nous référer (2) nous avons essayé de ca-

---

(1) March (L.), « Comparaison numérique des courbes statistiques ». (*Journal de la Société de Statistique de Paris*, 1905, pp. 261-262.)

(2) Julin (Arm.), « The economic progress of Belgium ». (*Journal of the Royal Statistic Society*, London, 1910-11, p. 251.)

ractériser l'ampleur de certains phénomènes économiques au moyen d'une méthode spéciale d'index-numbers que l'on trouvera décrite en détail dans le mémoire cité.

Les indices concernant la production, au nombre de sept, se rapportaient à la production des mines de houille, des carrières, à la fabrication du fer et de la fonte, à la fabrication de l'acier, à celle du zinc, du plomb et de l'argent, au nombre de moteurs à vapeur et à la force, en chevaux-vapeur, de ces moteurs. L'expression synthétique de ces indices, de 1884 à 1908, se trouve indiquée dans la première colonne du tableau suivant.

Le matériel statistique étant plus riche en ce qui concerne les échanges, nous avons pu considérer ici quinze indices, tels que les importations et exportations (commerce spécial), le tonnage de la navigation maritime, les transports de charbon et de coke par canaux, le nombre de voyageurs sur les chemins de fer de l'Etat et des compagnies, le trafic des marchandises par voie ferrée, etc. Le lecteur trouvera la moyenne de ces indices à la seconde colonne du tableau ci-après. Calculant ensuite l'augmentation ou la diminution de chaque indice synthétique de la première année à la seconde, de la seconde à la troisième et ainsi de suite, nous avons porté ces données dans les colonnes 3 et 4 de notre tableau. Le sens de ces variations, c'est-à-dire leur signe positif ou négatif est indiqué aux colonnes 5 et 6. Enfin, la colonne 7 est réservée au produit algébrique des signes, et les signes $+$ et $-$ indiquent que la concordance des variations est positive ou négative. Ensuite, utilisant la formule

$$\frac{c - d}{c + d} \qquad (54)$$

puisque aucun produit n'est nul, nous avons obtenu une fraction que nous avons réduite, pour la facilité du lecteur, en fraction décimale.

D'après ces explications, le tableau dont il s'agit présente les résultats suivants :

Exemple 1. — **Variation des indices de la production et des échanges en Belgique, de 1884 à 1908.**

ANNÉE	Indice de la production	Indice des échanges	Variation de l'indice de la production	Variation de l'indice des échanges	Sens des variations de la production	Sens des variations des échanges	Produit des signes
1884	100	100	7	1	+	+	+
1885	93	99	1	1	+	—	—
1886	92	100	— 8	— 4	—	—	+
1887	100	104	5	2	—	—	+
1888	105	106	10	4	—	—	+
1889	115	110	10	6	—	—	+
1890	125	116	5	4	+	—	—
1891	120	120	3	2	+	+	+
1892	117	118	2	1	+	—	—
1893	115	119	5	3	—	—	+
1894	120	122	2	3	—	—	+
1895	122	125	19	1	—	—	+
1896	141	126	9	6	—	—	+
1897	150	132	13	9	—	—	+
1898	163	141	— 23	— 3	—	—	+
1899	186	144	7	5	—	—	+
1900	193	149	30	2	+	—	—
1901	163	151	16	5	—	—	+
1902	179	156	19	6	—	—	+
1903	198	162	13	4	—	—	+
1904	211	166	10	9	—	—	+
1905	221	175	26	8	—	—	+
1906	247	183	12	4	—	—	+
1907	259	187	38	6	+	+	+
1908	221	181					

$$\frac{c - d}{c + d} = \frac{20 - 4}{24} = 0.66$$

Ainsi que nous l'avons dit plus haut, la concordance absolue de toutes les variations aurait pour expression numérique l'unité. L'expression obtenue au moyen des calculs qui précèdent est 0.66; elle signifie donc qu'il y a entre les échanges et la production, une même année, une dépendance marquée.

314. Il peut être intéressant de rechercher si cette dépendance est aussi accentuée lorsque, au lieu de considérer la même année de part et d'autre, on fait la somme algébrique des signes d'une année et de l'année suivante; par exemple, les échanges en 1884 avec la production en 1885, et ainsi de suite. On remarquera qu'alors le nombre des intervalles diminue d'une unité. Le lecteur pourra très facilement faire le contrôle du calcul en se servant du tableau qui précède. Le résultat numérique de cette recherche est :

$$\frac{19 - 4}{23} = \frac{15}{23} = 0,652$$

c'est-à-dire que la concordance des mouvements est, somme toute, la même que si l'on compare les données de la même année, et qu'elle reste nettement accusée.

On pourrait aussi se demander quels seraient les résultats obtenus si l'on considérait le rapport de la production la première année avec les échanges la seconde année. Le calcul donne :

$$\frac{17 - 6}{23} = \frac{11}{23} = 0,478$$

c'est-à-dire un indice de dépendance inférieur à une demi-unité, sensiblement moins élevé que les précédents. On peut en conclure que les échanges augmentent avant la production, c'est-à-dire que le commerce se développe d'abord sous l'influence d'une demande plus forte et que la production suit la même direction pour satisfaire cette demande. On a

d'ailleurs observé depuis longtemps que les importations devancent les exportations pendant les périodes d'expansion commerciale.

315. La méthode résumée ci-dessus est intéresante en elle-même et elle est importante à raison du point de départ qu'elle fournit à des méthodes plus compliquées que nous analyserons ensuite. Ces raisons seront suffisantes aux yeux du lecteur pour justifier une seconde application.

Nous en trouvons les données dans la statistique annuelle dressée par l'administration des mines en Belgique. Depuis 1885, cette administration, indépendamment du salaire moyen général des ouvriers de charbonnages, donne le salaire journalier moyen de l'ouvrier du fond; cette dernière valeur est plus homogène que la moyenne générale et par conséquent, elle est plus représentative; la période de 1885 à 1915 est d'ailleurs suffisante pour apprécier la valeur des changements.

D'autre part, le bénéfice moyen à la tonne est indiqué dans les statistiques des mines depuis de longues années. Existe-t-il une relation entre ces deux éléments? Et le taux du salaire des ouvriers producteurs, les ouvriers du fond, marque-t-il une dépendance quelconque avec le profit moyen à la tonne réalisé par les sociétés charbonnières? Dans le but d'apprécier avec exactitude la relation existante, on emploiera la méthode décrite précédemment; le tableau ci-contre donne les résultats du calcul.

Exemple 2. — **Bénéfice à la tonne et salaire journalier moyen de l'ouvrier du fond dans les charbonnages en Belgique, de 1885 à 1915.**

ANNÉES	Bénéfice à la tonne	Salaire journalier moyen de l'ouvrier du fond	Variation du bénéfice à la tonne	Variation du salaire journalier moyen	Sens des variations du bénéfice à la tonne	Sens des variations du salaire journalier moyen	Produit des signes
	Francs	Francs	Francs	Francs			
1885	0.40	3.20	0.10	0.26	+	+	+
1886	0.30	2.94	0.18	0.05	—	—	+
1887	0.48	2.99	0.17	0.11	—	—	+
1888	0.65	3.10	0.46	0.32	—	—	+
1889	1.11	3.42	1.73	0.78	—	—	+
1890	2.84	4.20	1.02	0.02	+	+	+
1891	1.82	4.18	1.22	0.62	+	+	+
1892	0.60	3.56	0.27	0.27	+	+	+
1893	0.33	3.29	0.07	0.04	—	+	—
1894	0.40	3.25	0	0.18	0	—	0
1895	0.40	3.43	0.12	0.06	—	—	+
1896	0.52	3.49	0.38	0.23	—	—	+
1897	0.90	3.72	0.15	0.22	—	—	+
1898	1.05	3.94	0.66	0.43	—	—	+
1899	1.71	4.37	2.55	0.84	—	—	+
1900	4.26	5.21	1.93	0.52	+	+	+
1901	2.33	4.69	0.92	0.30	+	+	+
1902	1.41	4.39	0.18	0.01	+	+	+
1903	1.23	4.38	0.48	0.18	+	+	+
1904	0.75	4.20	0.08	0.08	—	—	+
1905	0.83	4.28	1.08	0.70	—	—	+
1906	1.91	4.98	0.25	0.54	—	—	+
1907	2.16	5.52	0.79	0.35	+	+	+
1908	1.37	5.17	0.63	0.53	+	+	+
1909	0.74	4.64	0.24	0.21	+	—	—
1910	0.50	4.85	0.36	0.11	+	—	—
1911	0.14	4.96	0.20	0.38	—	—	+
1912	0.34	5.34	0.49	0.42	—	—	+
1913	0.83	5.76	0.20	0.58	+	+	+
1914	0.63	5.18	0.12	0.67	—	+	—
1915	0.75	4.51	—	—			

$$\frac{c-d}{n} = \frac{25-4}{29} = 0.724$$

Le résultat 0.724 indique qu'il existe une dépendance marquée entre le taux des salaires des ouvriers du fond et le bénéfice à la tonne réalisé par les charbonnages.

Cette dépendance se traduit dans l'année même où le bénéfice est réalisé, car si l'on unit la première année, pour les bénéfices, avec la seconde, pour les salaires, on n'obtient qu'un indice de dépendance très restreint :

$$\frac{c-d}{n} = \frac{18-10}{28} = 0,285$$

Ces exemples suffisent à montrer quels sont les services que l'indice de dépendance est capable de rendre aux chercheurs. L'indice présente l'avantage d'une grande simplicité, mais à raison de ce caractère même des opérations, il manque de précision, chose que le lecteur aura déjà remarquée avant qu'on ne l'en avertisse.

III. — Coefficient de dépendance.

316. Pour établir l'indice de dépendance, on admet comme postulat initial que chaque discordance entre les deux valeurs comparées annule une concordance. Bien que l'indice calculé d'après cette base donne déjà une idée approchée de la relation des deux courbes, la méthode est néanmoins exposée à un grave reproche : celui de ne pas tenir compte de l'ampleur des variations. Une légère discordance est tenue pour égale à une forte concordance ; il y a dans cette hypothèse quelque chose de contraire à la réalité et qui nuit à l'exactitude du résultat. C'est pour cette raison que Fechner, après avoir indiqué la méthode de recherche de l'indice, a proposé ensuite un procédé plus parfait connu sous le nom de coefficient de dépendance. La forme la plus simple de ce coefficient consiste à transformer l'indice

simple de dépendance en accouplant par voie de multiplication les grandeurs des variations comparées qui sont de même signe. La formule est identique à celle de l'indice simple, sauf que les lettres minuscules sont remplacées par des majuscules :

$$I = \frac{C - D}{C + D} \qquad (55)$$

Il est à remarquer que la formule du coefficient de dépendance proposée par Fechner, porte, tout comme l'indice de dépendance, sur les courbes réelles qui résultent des observations. Or, ces courbes sont difficilement comparables, car elles résultent d'unités différentes. C'est pour remédier à cet inconvénient que différents auteurs ont proposé de ramener les nombres de chaque série à leur valeur moyenne ou à une quantité invariablement liée à celle-ci; telle est la quantité qui résulte du rapport du taux de chaque année au taux moyen. La formule, dont le calcul arithmétique exige seulement la transformation des données originales en pour cent, est de la forme :

$$\frac{t \times 100}{m} = tm \qquad (56)$$

dans laquelle t représente le taux initial (valeur correspondante à l'année) ; m la moyenne de ces valeurs ; tm le pourcentage cherché.

Faisons application de cette méthode à notre exemple 2 relatif au rapport de dépendance entre le salaire journalier moyen de l'ouvrier du fond et le bénéfice moyen à la tonne réalisé par les charbonnages exploités en Belgique de 1885 à 1915.

Dans ce but, nous établissons le tableau suivant (ex. 3) en y portant les données énumérées ci-après :

Après avoir inscrit les années considérées (1885-1915), nous reproduisons dans la colonne 2 les données relatives

au bénéfice moyen à la tonne, et dans la colonne 3 celles concernant le salaire de l'ouvrier du fond ; nous faisons la moyenne arithmétique de ces valeurs et nous l'inscrivons au bas de la colonne correspondante ; comparant ensuite les valeurs absolues de chaque année à la valeur moyenne de la série, d'après la formule (56), nous écrivons les résultats du calcul dans la colonne 4 pour le bénéfice à la tonne, dans la colonne 5 pour le salaire. Les colonnes 6 et 7 sont réservées à l'inscription des variations de ces pourcentages de chacune des deux données, en procédant comme nous l'avons fait à l'exemple 2, de la première année à la seconde, de la seconde à la troisième et ainsi de suite.

Dans les trois colonnes suivantes (8, 9, 10) on note les produits des variations deux à deux en distinguant les produits positifs, négatifs et nuls et on fait la somme de chacune de ces variations. (*Voir tableau ci-contre.*)

Exemple 3. — **Bénéfice à la tonne et salaire journalier moyen de l'ouvrier du fond dans les charbonnages, en Belgique, de 1885 à 1915.**

Coefficient I *de dépendance.*

Années	Bénéfice à la tonne francs	Salaire journalier moyen d'ouvriers du fond francs	Rapport des données annuelles à la moyenne (pour unité)		Variations annuelles		Produits des variations la même année		
			Bénéfice %	Salaire %	Bénéfice %	Salaire %	Produits positifs	Produits négatifs	Produits nuls
	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1885	0 40	3 20	21 39	75.65	+ 5 35	+ 16 14	86 35	»	»
1886	0.30	2.94	16 04	69 51	− 9 63	− 1 34	12 89	»	»
1887	0 48	2.99	25.57	70.85	− 9 09	− 2 44	22.18	»	»
1888	0.65	3 10	34 76	73 29	− 24 60	− 7 56	185 98	»	»
1889	1 11	3 42	59.36	80 85	92 52	− 18 44	1706 07	»	»
1890	2.84	4.20	151 88	99 29	+ 54 55	+ 0 47	25.64	»	»
1891	1 82	4.18	97 33	98 82	+ 65 24	+ 14 66	956 42	»	»
1892	0.60	3 56	32 09	84 16	+ 14 44	+ 6 38	92.12	»	»
1893	0 33	3 29	17 65	77 78	− 3 74	+ 0 95	»	3 55	»
1894	0.40	3 25	21.39	76 83	0	− 4 26	»	»	0
1895	0.40	3 43	21 39	81 09	− 6 42	− 1 41	9 05	»	»
1896	0 52	3 49	27 81	82 50	- 20 32	− 5 44	110 54	»	»
1897	0 90	3.72	48 13	87 94	8 02	− 5.20	41.70	»	»
1898	1.05	3 94	56 15	93 14	− 35 30	− 10 16	358 65	»	»
1899	1 71	4 37	91.45	103 30	136 37	− 20 08	2738 31	»	»
1900	4.26	5.21	227 82	123.38	+103 22	+ 12 51	1291 28	»	»
1901	2.33	4 69	124 60	110 87	+ 49 20	+ 7 09	348 83	»	»
1902	1 41	4 39	75 40	103 78	+ 9 64	+ 0.23	2 22	»	»
1903	1 23	4 38	65 76	103 55	+ 25 65	+ 4 26	1091 27	»	»
1904	0.75	4.20	40 11	99.29	− 4 28	− 1 99	8.52	»	»
1905	0 83	4 28	44 39	101 28	57.75	− 16 45	949 98	»	»
1906	1 91	4 98	102 14	117 73	− 13 37	− 12 77	170 73	»	»
1907	2.16	5 52	115 51	130.50	+ 42 25	+ 8 28	349.83	»	»
1908	1.37	5 17	73 26	122 22	+ 33 69	+ 12.52	421.80	»	»
1909	0 74	4 64	39 57	109 70	+ 12 83	− 4 97	»	63 77	»
1910	0 50	4 85	26 74	114 67	+ 19 25	− 2 60	»	50 05	»
1911	0 14	4 96	7 49	117.27	− 10 69	− 8.97	95 89	»	»
1912	0 34	5 34	18.18	126 24	26 21	− 9 94	260.53	»	»
1913	0 83	5 76	44.39	136.18	+ 11.23	+ 13 71	153 96	»	»
1914	0 63	5 18	33.16	122 47	− 6.95	+ 15.84	»	110.09	»
1915	0.75	4.51	40 11	106.63	»	»	»	»	»
	M = 1.87	M = 4.23					11,490.74	227.46	

$$I = \frac{C - D}{N} = \frac{11{,}490.74 - 227.46}{11{,}718.20} = 0.96$$

Dans le premier cas (indice de dépendance, *i*) nous avons obtenu *le nombre de variations de même sens* entre le bénéfice à la tonne réalisé par les charbonnages et le salaire journalier moyen attribué aux ouvriers du fond. Nous pouvons conclure, en présence des résultats, que les variations en question sont souvent concomitantes : sur 100 variations, il y en a 72 qui sont de même sens. C'est plus qu'il n'en faut pour pouvoir affirmer que souvent le salaire augmente quand le bénéfice s'accroît.

Mais on peut se poser la question suivante : le salaire augmente-t-il dans la même proportion que les bénéfices ? Le coefficient de dépendance I donné par la formule (55) (Cfr. n° 314) permet de dire qu'il en est presque toujours ainsi : Le lien de dépendance entre les deux phénomènes est donc bien plus accentué qu'il ne paraissait l'être par le simple indice *i* : en effet, l'union qui existe entre les proportions des variations est très accentuée si nous considérons qu'il est marqué par une expression numérique d'un ordre élevé : 0.96. Le coefficient I a donc confirmé et précisé celui qui précède.

Ce résultat intéressant est à noter et nous passons ensuite à une expression différente, l'expression J, dans laquelle on substitue à la somme des valeurs absolues des produits : C + D la moitié de la somme des carrés des variations comparées. Le but de cette substitution est de permettre l'introduction du coefficient de covariation dans les calculs algébriques.

317. L'expression J peut s'écrire :

$$J = \frac{C - D}{\frac{\Sigma v^2 + \Sigma v_1^2}{2}} \quad \text{ou mieux} \quad \frac{\Sigma v v_1}{{}^1/_2 (\Sigma v^2 + \Sigma v_1^2)} \qquad (57)$$

Le calcul de l'expression $\Sigma v v_1$, somme égale à C — D est contenu dans les colonnes 8 et 9 du tableau précédent (exemple 3). Les produits positifs égalent + 11,490.74 ; les

produits négatifs valent — 227.46. Leur somme algébrique est donc égale à + 11,263.28, valeur qui formera le numérateur de la fraction.

Le dénominateur de la fraction est formé de la moyenne des sommes des carrés des deux séries de variations : v = les variations de la première, v_1 celles de la seconde. Les données à porter au carré se trouvent inscrites dans les colonnes 6 et 7 du tableau dont il vient d'être parlé. Nous avons $\Sigma v^2 = 58{,}242.22$ et $\Sigma v_1^2 = 5{,}133.97$ dont la moyenne est : 31,688.09. En remplaçant les expressions littérales de la formule (57) par les données numériques ci-dessus indiquées, nous avons :

$$\frac{\Sigma v v_1}{{}^1/_2 \left[\Sigma v^2 + \Sigma v_1{}^2\right]} = + \frac{11{,}263.28}{{}^1/_2 31{,}688.09} = + 0.70 \text{ environ}$$

Le coefficient est égal à l'unité lorsque les modifications que subissent les deux séries dans leur évolution chronologique, sont de même sens *et égales;* il diminue progressivement dans le cas contraire. L'importance du coefficient J est, dans l'espèce, démontrée par le fait que l'indice de Fechner, en outre de l'union existant entre les proportions de variations, atteignait 0.96, tandis que J marquait 0.70 seulement. C'est que les variations des bénéfices des charbonnages sont beaucoup plus larges que celles observées dans les salaires; bien que le salaire de l'ouvrier du fond soit sensible aux variations dans les bénéfices, et cela la même année, les modifications qu'il subit sont bien moindres que celles du profit à la tonne.

318. Si, au dénominateur, on remplace par la moyenne géométrique la moyenne arithmétique employée pour le calcul du coefficient J, on obtient un coefficient nouveau, que M. March a appelé coefficient K et qui est de la forme :

$$K = \frac{\Sigma v v^1}{\sqrt{\Sigma v^2 \cdot \Sigma v_1{}^2}} \qquad (58)$$

dans laquelle le numérateur représente la somme des produits deux à deux de chacune des séries et le dénominateur la racine carrée du produit des sommes des carrés.

D'après M. March, le coefficient K semblerait devoir être préféré au coefficient J. Les raisons données par notre savant collègue sont à retenir. Pour passer au coefficient K, on change les unités de mesure des v et des v_1. On prend comme unité de mesure des v, $\sqrt{\Sigma v^2}$ et comme unité de mesure des v_1, $\sqrt{\Sigma v_1^2}$. Remplaçant dans l'expression de J, les v et v_1 par ces nouvelles expressions, on a :

$$J = \frac{\Sigma \frac{v}{\sqrt{\Sigma v^2}} \times \frac{v_1}{\sqrt{\Sigma v_1^2}}}{1/2\left[\frac{\Sigma v^2}{\Sigma v^2}\right] + \left[\frac{\Sigma v_1^2}{\Sigma v_1^2}\right]} = \frac{\Sigma v v_1}{\sqrt{\Sigma v^2}.\sqrt{\Sigma v_1^2}} = r$$

dans laquelle $r = 1$ quand v et v_1 sont dans un rapport constant, car si l'on a constamment $v = kv_1$

$$r = \frac{\Sigma k v_1^2}{\sqrt{\Sigma k^2 v_1^2}\sqrt{\Sigma v_1^2}} = \frac{\Sigma v_1^2}{\Sigma v_1^2} = 1$$

Lorsque les sommes Σv_2 et Σv_1^2 ne sont pas très différentes, les deux coefficients ne s'écartent pas beaucoup l'un de l'autre. Mais à mesure que les sommes Σv_2 et Σv_1^2 se différencient, « leur moyenne géométrique s'écarte de plus en plus de leur moyenne arithmétique, dans les conditions où une parabole s'écarte de sa tangente (1) ». Le coefficient K tend, dans ce cas, vers une valeur fixe, moins petite que celle du coefficient J. En prenant une valeur très réduite, J exprime bien la fixité relative d'une série par rapport à l'autre, ce que l'on pourrait à première vue considérer comme un signe d'indépendance; mais cet aspect des choses n'est pas complet, car si même les variations d'une série sont très petites, par rapport aux oscillations de l'autre, s'il y a entre ces changements un synchronisme

(1) MARCH (L.), Cfr. *op. cit.*, *loc. cit.*, p. 268.

parfait, on ne pourra s'empêcher de conclure qu'un lien étroit de dépendance existe entre les deux phénomènes. L'exemple 3 convient parfaitement pour l'application du coefficient K, car nous avons vu que $\Sigma\ v^2$ est très différent de $\Sigma\ v_1^2$ et que J n'a qu'une valeur plus faible, tandis que I, qui exprime mieux le synchronisme des variations, a une valeur élevée.

Le numérateur de la fraction d'où se tire le coefficient K est formé de la somme algébrique des variations positives et négatives prises deux à deux. Les données numériques se trouvent inscrites aux colonnes 8 et 9 du tableau relatif à l'exemple 3. Quant au dénominateur, il est formé de la racine carrée du produit de $\Sigma\ v^2$ et $\Sigma\ v_1^2$ dont la valeur numérique est respectivement 58,242.22 $\times$ 5,133.97 ; pour simplifier les calculs, nous supprimons les deux décimales, en forçant le dernier chiffre du premier de ces nombres.

$$K = \frac{\Sigma\ v\ v_1}{\sqrt{\Sigma\ v^2 \times \Sigma\ v_1^2}} = \frac{11.263}{\sqrt{58.242 \times 5134}} = \frac{11.263}{17.292} = +\ 0.65$$

Le coefficient K indique par conséquent une liaison fort marquée entre les deux phénomènes considérés et permet d'interpréter définitivement les données numériques des séries. En nous rapportant à l'indice I nous voyons que la concomitance des hausses et des baisses se présente souvent ; par le coefficient J, nous constatons que, tandis que l'une des séries présente de grandes oscillations, la seconde n'en ressent que de beaucoup moindres ; enfin, le coefficient K rectifie ce qu'il pourrait y avoir d'erroné dans nos conclusions basées sur I et sur J, en nous avertissant qu'à un mouvement dans la première série correspond souvent un mouvement dans la seconde ; bien que de moindre ampleur, le mouvement dont il s'agit n'en est pas moins un indice sérieux de la liaison existant entre les deux phénomènes.

Les applications qui précèdent montrent clairement l'intérêt qui s'attache à la détermination du coefficient de dépendance; elles n'exigent pas, en général, de calculs trop longs ou difficiles. La dernière forme, celle du coefficient K, nous conduit directement à celle que les Anglais ont appelée coefficient de corrélation et qu'ils désignent sous le symbole r.

IV. — Coefficient de covariation.
(Coefficient de corrélation (r de Pearson).

319. Pour nous accoutumer à l'idée de la covariation, représentons-nous mentalement deux courbes chronologiques, de telle nature qu'à toute variation de l'une corresponde, au même moment, une variation égale de l'autre. Ces variations peuvent être égales comme grandeurs absolues ou l'être seulement comme grandeurs proportionnelles. Maintenant que le lecteur est familier avec l'idée de moyenne, il saisira immédiatement que les écarts ou déviations autour de la moyenne de chacune de ces courbes sont égaux entre eux; représentons-nous la moyenne par une droite tracée à travers la courbe; d'après l'hypothèse qui précède, nous comprenons fort bien que tout écart au-dessus ou au-dessous de la courbe A correspond exactement à un écart égal au-dessus ou au-dessous de la courbe B. Désignons par $x_1, x_2, x_3 \ldots x_n$ les écarts successifs de A et par y_1, $y_2, y_3 \ldots y_n$ les écarts de B; les sommes algébriques des x_n et des y_n seront égales entre elles et leur produit formera la quantité $\Sigma\,(x_n\,y_n)$. D'autre part, le produit des *standard deviations* (σ) des écarts, multiplié par le nombre de termes sera égal au produit $\Sigma\,(x_n\,y_n)$ de sorte que le coefficient de covariation s'exprimera par l'unité :

$$(x_n\,y_n) = n \cdot \sigma_x \cdot \sigma_y = 1 \qquad (1)$$

Le coefficient de covariation oscille entre les deux limites

(1) On en trouvera la démonstration mathématique dans BRESCIANI : « Sui metodi per la misura delle correlazioni (*Giornale degli economisti*, vol. XXXV, 1909, t. I, p. 492, note) et dans BOWLEY, *Elements of statistics*.

0 et 1 ; il sera donc d'autant plus élevé qu'il se rapprochera de l'unité et d'autant plus faible qu'il s'en écartera davantage. Lorsque le coefficient est affecté du signe +, il y a corrélation directe entre les faits envisagés ; accompagné du signe — il signifie que la corrélation entre les faits est inverse.

320. Au lieu d'une courbe, envisageons maintenant les éléments à comparer sous la forme d'un tableau statistique. Disposons les variables en séries en les groupant comme dans une table à double entrée. Les tables de cette espèce contiennent un nombre de colonnes égal au nombre de classes admises pour un phénomène donné, en fonction d'un autre. Elles sont très fréquentes en statistique, les auteurs qui ont traité de la covariation en donnent de nombreux exemples. On remarquera qu'il n'est pas nécessaire que le partage par classes soit identique, pourvu que les classes aient un intervalle égal, pour chacune des deux variables ; néanmoins, le partage par classes comprises entre les mêmes limites est la règle habituelle, comme dans l'exemple ci-après dont nous tirons les éléments de la « Statistique du mouvement de la population et de l'état civil en Belgique en 1890 » :

Ex. 4 **Ages des époux au moment de la célébration du mariage.**

Ages des Hommes	Ages des Femmes							
	— 21	21 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50	Total
— 21	774	426	128	40	26	2	1	1,397
21 - 25	3,401	6,086	2,157	409	96	39	25	12,213
25 - 30	2,073	6,739	5,720	1,343	382	133	49	16,439
30 - 35	498	1,669	2,646	1,601	442	180	78	7,114
35 - 40	140	441	838	852	512	221	105	3,109
40 - 45	40	170	344	380	291	250	133	1,608
45 - 50	21	54	134	195	187	208	162	961
Totaux...	6,947	15,585	11,967	4,820	1,936	1,033	553	42,841

Dans l'exemple qui précède, nous avons dû négliger les mariages contractés à partir de 50 ans, parce que le matériel statistique dont nous disposions ne donne que le nombre de mariages de 10 en 10 ans, au delà de 50 ans, de telle sorte que les intervalles de classes n'auraient pas été homogènes ; ceci est un nouvel exemple de l'erreur dans laquelle on tombe fréquemment en ne publiant pas les données statistiques sous une forme complète, l'effet de cette omission est de mettre obstacle à l'application des procédés mathématiques. Néanmoins le tableau même limité reste très intéressant et très démonstratif puisqu'il comprend près de 43,000 mariages sur un peu plus de 44,000.

Le tableau qui précède réunit les éléments nécessaires à l'étude des relations corrélatives entre l'âge des femmes et celui des maris, pour les mariages contractés en 1890, en Belgique, entre des personnes âgées de moins de 50 ans. Les classes d'âges des femmes se trouvent inscrites en tête des colonnes verticales, celles des hommes sont portées à la gauche de colonnes horizontales, que nous appellerons « lignes » pour les distinguer des premières. Les nombres qui figurent dans le corps de la table et qui constituent chacun un des points de rencontre d'une colonne verticale avec une ligne horizontale indiquent pour un âge donné des femmes et pour un âge donné des maris, le nombre de couples qui ont contracté mariage cette année.

La colonne verticale 2, par exemple, nous apprend que parmi les femmes de 21 à moins de 25 ans, qui se sont mariées en 1890 et dont le nombre total s'élève à 15,585, 6,739 ont épousé des hommes d'un âge égal au leur (dans les limites de la classe 21-25) ; 170 ont épousé des hommes de 40 à moins de 45 ans, 54 se sont mariées à des hommes de 45 à moins de 50 ans. La première « ligne » horizontale nous montre que sur 42,841 mariés, 1,397 étaient âgés de moins de 21 ans ; le plus grand nombre ont épousé des filles de leur âge, mais quelques-uns ont pris en mariage des

femmes beaucoup plus âgées qu'eux; 26, par exemple, ont épousé des femmes de 35 à 40 ans.

Les tables de corrélation fournissent ainsi des éléments très curieux utilisables au moyen d'une simple lecture, mais elles ne donnent pas la réponse précise qui est à rechercher par le moyen du coefficient de corrélation. D'autre part, les tables de corrélation supposent qu'il s'agit de deux variables partagées en classes de fréquence. Lorsque la corrélation à rechercher vise des séries chronologiques composées d'une donnée unique par année, les tables de corrélation ne trouvent pas leur application; on se borne dans ce cas à confronter les deux données relatives à la même année.

321. La théorie du calcul des covariations a été exposée avec une grande habileté par les statisticiens-mathématiciens anglais, notamment par MM. Karl Pearson et Udny Yule. Nous essayerons d'en résumer ci-après l'essentiel en donnant à notre exposé la forme la plus simple et la plus accessible aux lecteurs qui ne sont pas familiarisés avec l'emploi des formules mathématiques.

La théorie des covariations est basée sur la proposition suivante : lorsque deux variables sont absolument indépendantes, les distributions de fréquence dans les colonnes verticales et les lignes horizontales d'une table de contingences sont similaires et la distribution dans chaque colonne est semblable à celle de la colonne des totaux (1) ; si l'on attribue à chacune de ces distributions un axe spécial, X ou Y, les moyennes des séries doivent se trouver sur les lignes verticales et horizontales M_1 M; M_2 M, comme dans la figure 26 (2) :

Dans le diagramme ci-après OX et OY sont les

(1) Pour la démonstration de cette propriété, cfr. Udny YULE, « An introduction to the theory of statistics », ch. V (*Manifold classification*).

(2) Nous adoptons comme base de la démonstration la figure employée par Udny YULE, *loc. cit.*, ch. IX (correlation), p. 168, et nous employons la notation de notre éminent collègue.

échelles des deux variables; M_1 est donc la moyenne de la valeur de X et M_2 la moyenne de la valeur de Y; en d'autres termes, M_1 est la moyenne de la valeur des abscisses et M_2 celle des ordonnées. La figure 26 est le type théorique de représentation de deux variables dans l'éventualité d'une indépendance absolue de ces deux variables. Dans la pratique ce cas ne se rencontre pas, parce que l'absence de tout rapport de dépendance n'est jamais absolument complète. Quoi qu'il en soit, cette conception purement idéale d'une indépendance absolue des variables a son utilité en ce sens qu'elle sert de base de comparaison dans le calcul des relations corrélatives qui se manifestent entre les deux variables considérées.

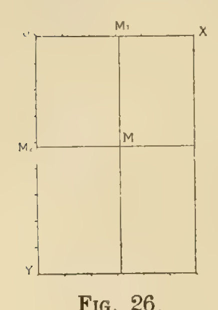


Fig. 26.

En effet, hors le cas d'indépendance absolue, les distributions de fréquence ne sont plus similaires et les moyennes des séries ne tombent plus en M_1 M et M_2 M, mais déterminent des courbes qui s'écartent plus ou moins de ces points. Or, ce sont les écarts constatés entre les positions théoriques et les positions réelles des variables qui sont à calculer en vue de la recherche du degré de covariation entre les phénomènes observés.

Partant de cette considération, si, conformément au dispositif de la figure 27 ci-contre, nous convenons de représenter dans les quadrants 2 et 4 les oscillations de M_1 y et M_2 x, nous pouvons admettre que ces points occupent une infinité de positions comprises entre les limites de ces quadrants.

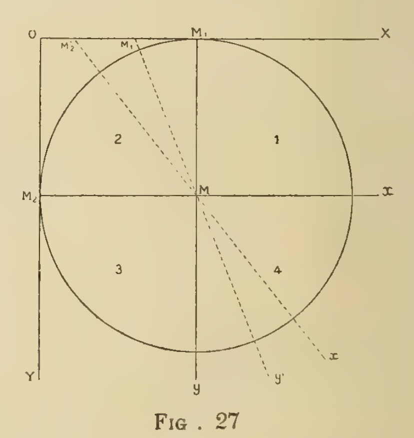


Fig. 27

Au plus la somme des écarts sera grande et au plus le degré de covariation entre les deux phénomènes sera élevé. Il deviendra parfait lorsque $M'_1\ y'$ se confondra avec $M'_2\ x'$. Dans ce cas, la somme des écarts sera maximum et sa valeur vaudra 90° parce que les arcs des angles x' M y et y' M x seront alors complémentaires.

Il résulte de ce qui précède que lorsque la somme des écarts est égale à zéro, il y a indépendance absolue entre les deux variables et que lorsque cette somme est maximum, la corrélation est parfaite entre les deux phénomènes ; toute position intermédiaire détermine un degré de corrélation plus ou moins affirmé.

322. Ceci établi, il reste à voir comment se calculent les écarts formés par les oscillations de $M'_1\ y'$ et $M'_2\ x'$ à partir de $M_1\ y$ et de $M_2\ x$.

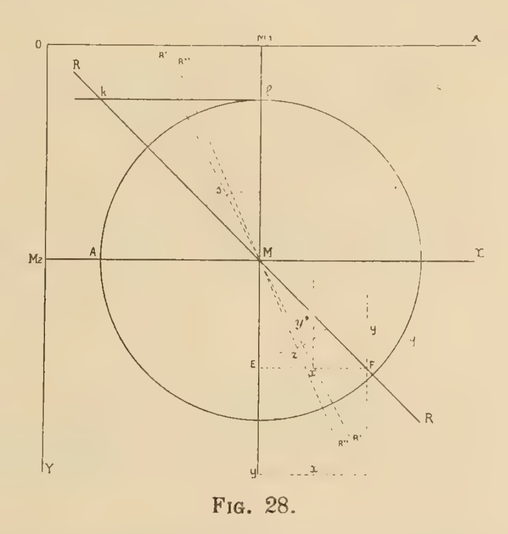


Fig. 28.

A cette fin, il y a notamment lieu de démontrer que quel que soit l'aspect sous lequel se présente la distribution de fréquence des deux variables, le point M déterminé par l'intersection de $M_1\ y$ et de $M_2\ x$ est le lieu géométrique de la moyenne de la distribution totale des fréquences des deux variables à la fois.

Mais avant d'aborder cette démonstration il est utile et même nécessaire d'étudier la genèse et la nature des équations qui forment la base du raisonnement.

Considérons la figure 28 ci-dessus :

Lorsque R R forme avec l'axe vertical $M_1\,y$ un angle de 45° nous avons :

$$K\,l = l\,M,$$

d'où $\frac{K\,l}{l\,M}$, rapport que M. Udny Yule désigne par $b_1 = 1$.

Si nous représentons par x et par y les écarts à partir de $M_1\,y$ et de $M_2\,x$, il suit de ce qui précède que $x = y$ et que le nombre de déviations en x est égal au nombre des observations du type y multiplié par b_1, ce qui donne :

$$\Sigma\,(x) = n\,b_1\,y \qquad (59)$$

égalité qui se vérifie pour toutes les positions de R R, ainsi que le montre l'analyse des cas suivants :

Premier cas. — Si nous faisons tourner R R dans le sens A l, il arrivera un moment où R R se trouvera en R' R' et b_1 aura alors pour valeur $\frac{b_1}{2}$ ou 1/2. Les déviations en x n'ont pas changé, mais les écarts ont été multipliés par 2. Il viendrait donc d'après la formule (59) :

$$\Sigma\,(x) = \frac{n\,b_1\,2y}{2}$$

et, en simplifiant, il vient comme dans la formule (59) :

$$\Sigma\,(x) = n\,b_1\,y$$

Généralisant la formule et désignant par p toute position quelconque de R R comprise entre les angles s et z (la position R'' R'' par exemple dans la figure 28), il vient :

$$\Sigma\,(x) = \frac{n\,b_1\,p\,y}{p}$$

ou en simplifiant :

$$\Sigma\,(x) = n\,b_1\,y$$

ce qui vérifie la formule indiquée plus haut.

Deuxième cas. — On obtiendrait le même résultat en considérant la valeur de l'écart x dans sa position initiale E F. Dans ce cas, lorsque R R coupe b_1 en $\frac{b_1}{2}$ l'écart x devient $\frac{x}{2}$ et l'écart y tombe en y' (*Cfr. fig. 28*). La formule se présente alors comme suit :

$$\frac{\Sigma\,(x)}{2} = \frac{n\ b_1\ y'}{2}$$

ou

$$\Sigma\, x = n\, b_1\, y'$$

ce qui vérifie la formule pour le deuxième cas.

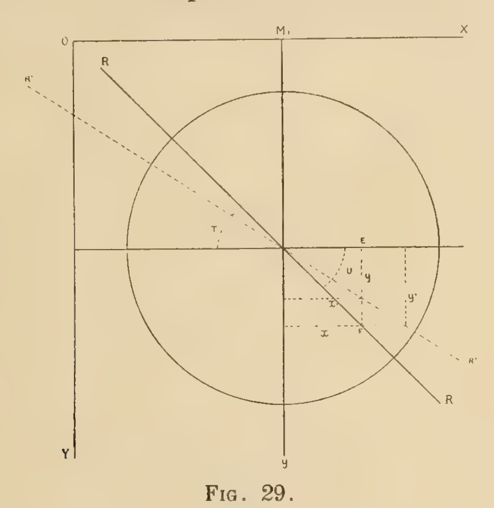


FIG. 29.

Troisième cas. — Pour toutes les positions de R R dans les limites des angles s et z (*Cfr. fig. 28*), le rapport $\frac{K}{l\,M}$ varie de 0 à 1.

Pour toutes les positions de R R dans les limites des angles T et U (*voir fig. 29*) le rapport $\frac{K\,l}{l\,M}$ varie de 1 à x.

Lorsque $\frac{K\,l}{l\,M} > 1$ soit $\frac{K\,l}{l\,M} = p\ b_1$, la formule générale devient :

$$p\ \Sigma\,(x) = n\, p\, b_1\, y$$

d'où par simplification :

$$\Sigma\,(x) = n\, b_1\, y$$

En considérant l'écart y dans sa position d'origine E F et en supposant que R R coupe b_1 en 2 b_1 l'écart y devient $\frac{y}{2}$ et la formule se représente alors sous la forme

$$\Sigma\,(x) = \frac{n\,2\,b_1\,y}{2} \text{ où } \Sigma\,(x) = n\,b_1\,y$$

Si, comme dans la figure 29, R' R' a une autre valeur que $\frac{b_1}{2}$, soit p, on se trouve dans le cas de la formule générale.

323. Les considérations qui précèdent concernant les trois cas relatifs aux valeurs de $\frac{\mathrm{K}\,l}{l\,\mathrm{M}}$, nous fournissent les éléments nécessaires à la compréhension complète de la démonstration du théorème initial.

Soit la figure ci-dessous :

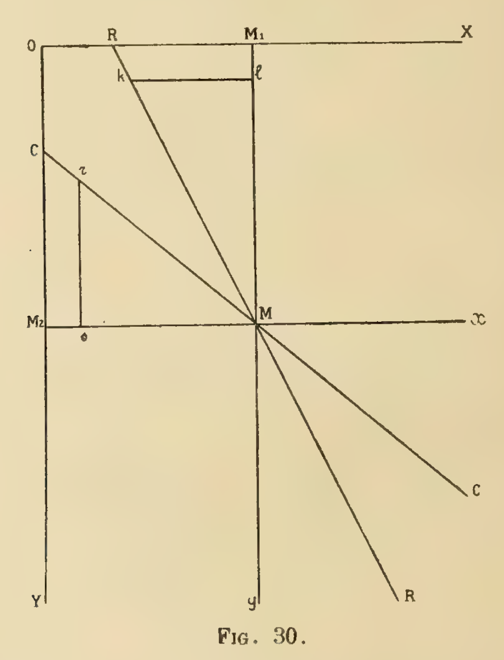


Fig. 30.

Considérons le cas dans lequel les moyennes des colonnes horizontales de la table de covariation ou de contingence tombent exactement sur une droite R R. Soit M_2 la valeur

moyenne de Y et R R coupant l'horizontale $M_2\,x$ au point M. Il y a lieu de démontrer que la verticale $M_1\,y$ qui coupe $M_2\,x$ en M doit couper O X en M_1 moyenne de l'axe des X.

Soit l'inclinaison de R R sur $M_1\,y$ exprimée par la tangente de l'angle M_1 M R ou le rapport $\frac{K\,l}{l\,M}$, désigné par b_1, et soient x et y les déviations à partir de M y et de M x.

Pour chaque colonne horizontale du type y, pour lequel le nombre d'observations est n, il vient $\Sigma\,(x) = n\,b_1\,y$. Par conséquent, pour toute la table, puisque $\Sigma\,(n\,y) = 0$, il viendra : $\Sigma\,(x) = b_1,\ \Sigma\,(n\,y) = 0$. M_1 doit donc être la moyenne de x et M devient la moyenne de toute la distribution.

Sachant que RR passe en M_1 il reste à déterminer b_1. Cette détermination comme le dit M. Udny Yule, peut s'établir dans les termes du produit moyen p de tous les couples de déviation associés x et y.

L'expression du produit moyen p, d'après la notation d'usage, s'écrit comme suit :

$$p = \frac{1}{N}\,\Sigma\,(xy) \qquad (60)$$

ce qui donne pour chaque colonne horizontale :

$$\Sigma\,(xy) = y\,\Sigma\,(x) = n\,b_1\,y^2 \quad [1]$$

Pour toute la table, il viendra :

$$\Sigma\,(xy) = b_1\,\Sigma\,(n\,y^2) = N\,b_1\,\sigma_y^2$$

Il s'agit d'obtenir la valeur de $b_1\,\sigma_y^2$.

D'après la formule ci-dessus, nous avons :

$$b_1\,\sigma_y^2 = \frac{\Sigma\,(x\,y)}{N}$$

(1) Nous avons vu que $\Sigma\,(x) = n\,b_1\,y$. En remplaçant $\Sigma\,(x)$ par sa valeur, nous avons : $y\,n\,b_1\,y = n\,b_1\,y^2$. Donc $y\,\Sigma\,(x) = n\,b_1\,y^2$.

En remplaçant par p le second membre de l'équation, nous aurons :

$$b_1 \sigma_y^2 = p$$

Enfin, en tirant la valeur de b_1 dans la formule, il viendra :

$$b_1 = \frac{p}{\sigma_y^2} \qquad (61)$$

De même b_2, c'est-à-dire le rapport $\frac{r\,s}{s\,M}$ (*Cfr. fig. 30*) mesurant les écarts en C C serait déterminé comme suit :

$$b_2 = \frac{p}{\sigma_x^2} \qquad (62)$$

Il est d'usage d'écrire ces deux équations en se servant d'une notation quelque peu différente. Voici cette expression :

$$r = \frac{p}{\sigma_x \sigma_y} \qquad (63)$$

324. Dès lors pour retrouver la valeur de b_1, en considérant la formule précédente, il suffit de faire subir à celle-ci les transformations suivantes :

Nous avons vu que $p = b_1 \sigma_y^2$. Par conséquent, en remplaçant p par sa valeur dans l'équation (63), nous obtiendrons :

$$r = \frac{b_1 \sigma_y^2}{\sigma_x \sigma_y} \qquad (64)$$

En isolant b_1, la formule se transforme comme suit :

$$r = b_1 \frac{\sigma_y^2}{\sigma_x \sigma_y}$$

Simplifiant la fraction de la formule, il vient :

$$r = b_1 \frac{\sigma_y}{\sigma_x}$$

d'où :

$$b_1 = r \frac{\sigma_x}{\sigma_y} \qquad (65)$$

On trouverait de même la valeur de b_2 :

$$b_2 = r \frac{\sigma_y}{\sigma_x} \qquad (66)$$

De ce qui précède, on peut déduire aisément que l'expression du coefficient de covariation est de la forme :

$$r = \frac{p}{\sigma_x \sigma_y} \qquad (67)$$

En effet, r est fonction de b_1 et de b_2 en ce sens que quand ces deux variables sont absolument indépendantes et égales à 0, $r = 0$.

Or, nous avons pu remarquer en recherchant la formule de détermination de b_1 et de b_2 que b_1 et b_2 sont des valeurs implicitement comprises dans le second membre de l'équation :

$$r = \frac{p}{\sigma_x \sigma_y}$$

Notons également que le signe de r est celui du produit moyen p et que par conséquent r est positif lorsqu'à de grandes valeurs de x correspondent de grandes valeurs de y et qu'il est négatif dans le cas opposé.

Enfin, la valeur numérique de r a pour limites $\pm$ 1. Si $r = \pm 1$, il s'ensuit que toutes les variations accouplées sont entre elles dans le rapport $\frac{x}{y} = \frac{\sigma_1}{\sigma_2}$.

Ces diverses propriétés de r se vérifiant toutes par l'équation de la forme

$$r = p / \sigma_x \sigma_y,$$

nous pouvons conclure que l'équation précitée est l'expression du degré des rapports de covariation entre deux variables et que r est le coefficient de covariation de deux variables données.

325. Pour le calcul numérique du coefficient r, trois cas sont à envisager :

A. — Les deux variables, disposées dans l'ordre chronologique, sont comparées entre elles de façon à saisir la relation qu'elles présentent sur toute l'étendue de temps considérée.

B. — Au lieu d'envisager uniquement cette relation, on étudie et on compare les variables de façon à mettre en évidence les variations portant sur une période de temps assez courte ; il est évident que le rapport de deux faits l'un à l'autre peut être différent quand il s'applique à un nombre d'années restreint et quand il concerne des variations séculaires.

C. — Le coefficient r appliqué à des variables groupées par classe ne se peut calculer de la même manière que pour des séries chronologiques. Une méthode spéciale s'impose pour une telle recherche.

Nous examinerons successivement ces trois cas auxquels correspond respectivement une méthode particulière et nous donnerons chaque fois un ou plusieurs exemples appropriés. Les calculs à effectuer, quoique assez laborieux, n'offrent point de difficultés spéciales. Quant à la formule générale du coefficient r, elle ne subit pas de modification essentielle : dans le cas B, la moyenne générale est

remplacée par une moyenne plus courte et les écarts se notent de la donnée primitive à la donnée nouvelle ; dans le cas C, la recherche de l'intervalle de classe joue un rôle important et la recherche des écarts se fait à l'aide d'un tableau spécial. Ces particularités n'altèrent point la formule générale, que l'on écrit :

$$r = \frac{\Sigma\ (xy)}{n\ \sigma_x \sigma_y} \qquad (68)$$

dans laquelle r désigne le coefficient cherché, $\Sigma\ (x\ y)$ la somme algébrique des produits des écarts à la moyenne, n le nombre d'intervalles ou de termes, σ_x la standard déviation de la première variable et σ_y la standard déviation de la seconde variable.

En représentant par P la moyenne de la somme des produits des écarts, la formule devient :

$$r = \frac{P}{\sigma_x\ \sigma_y} \qquad (69)$$

d'où il suit que le coefficient r est égal au quotient de la moyenne de la somme des produits des écarts par le produit des racines carrées de la moyenne des écarts des variables x et y et nous pouvons poser :

$$\frac{P}{\sigma_x\ \sigma_y} = \frac{\Sigma\ (xy)}{n\ \sigma_x \sigma_y} \qquad (70)$$

Le coefficient r est appliqué au calcul des variations de deux, de trois ou d'un plus grand nombre de variables. Nous envisagerons uniquement à cet endroit le cas de deux variables corrélatives, un paragraphe spécial sera réservé à l'examen et à l'application des méthodes lorsque trois variables sont considérées en même temps. Quant à l'étude des variables en plus grand nombre que trois, nous la passerons sous silence et nous renverrons le lecteur aux ouvrages des auteurs qui ont exposé ce point : nous avons

pris cette résolution à cause de la complication des méthodes quand il s'agit de plus de trois variables et parce que nous partageons l'avis de M. L. March qui pense que l'étude de la corrélation à deux variables est surtout celle qui doit retenir l'attention étant à peu près la seule qui ait été véritablement féconde (1).

326. Sous ce numéro nous étudierons l'application du coefficient r exclusivement à deux variables disposées selon l'ordre chronologique (cas A). (Cfr. n° 318.)

Un premier exemple nous est fourni par le tableau des lectures hebdomadaires de deux thermomètres souterrains t_1 et t_2 faites à l'observatoire d'Edimbourg et rapportées par M. Thomas Heath, astronome adjoint à cette Institution.

Dans la première colonne du tableau nous portons les dates des observations faites; pour la facilité du lecteur nous nous sommes borné aux trois premiers mois de l'année (1899); les colonnes 2 et 3 sont réservées à l'inscription des indications des thermomètres relevées chaque semaine, à la date en regard. La moyenne arithmétique simple de ces données est inscrite au bas de la colonne. Les colonnes 4 et 5 contiennent les écarts, avec leur signe, de chaque observation à la moyenne; sous les numéros 6 et 7 on trouve les carrés de ces écarts; enfin, la colonne 8 donne le produit des écarts entre eux, avec leur signe. On fait la somme des x^2 et des y^2, on en prend la moyenne et on en extrait la racine carrée. Le numérateur de la fraction est le produit Σxy; on ne doit pas oublier qu'il s'agit d'une somme algébrique. Le dénominateur se compose du produit du nombre d'observations par σ_x et σ_y.

(1) MARCH (L.), « Essai sur un mode d'exposer les principaux éléments de la théorie statistique ». (*Journal de la Société de Statistique de Paris*, 1910, p. 485.)

En se reportant au tableau suivant, le lecteur suivra avec la plus grande facilité la marche des calculs. Le résultat est positif et est égal à $r = 0{,}984$, c'est-à-dire qu'entre les indications fournies par les thermomètres t_1 et t_2 il y a une corrélation très étroite qui s'exprime par une donnée proche de l'unité.

Exemple 5.

Variations de deux thermomètres souterrains à Edimbourg.

(Matériel extrait des *Trans. Roy. Soc. of Edinburgh*, t. 40, 1900-1901, pp. 169-170, tab. I.)

Coefficient de Pearson (r)

DATES (1899)	Lectures hebdomadaires des thermomètres		Écarts à la moyenne de		x^2	y^2	xy
	t_1	t_2	t_1 (x)	t_2 (y)			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Janvier 2 . .	47.50	46.90	+0.39	+1.54	0.1521	2.3716	+0.6006
» 9 . .	47.49	46.72	+0.38	+1.36	0.1444	1.8496	+0.5168
» 16 . .	47.43	46.42	+0.32	+1.06	0.1024	1.1236	+0.3392
» 23 . .	47.3[illegible]	46.12	+0.25	+0.76	0.0625	0.5776	+0.1900
» 30 . .	47.32	45.83	+0.21	+0.47	0.0441	0.2209	+0.0987
Février 6 . .	47.22	45.52	+0.11	+0.16	0.0121	0.0256	+0.0176
» 13 . .	47.17	45.19	+0.06	−0.17	0.0036	0.0289	−0.0102
» 20 . .	47.08	44.86	−0.03	−0.50	0.0009	0.2500	+0.0150
» 27 . .	46.97	44.68	−0.14	−0.68	0.0196	0.4624	+0.0952
Mars 6 . .	46.88	44.54	−0.23	−0.82	0.0529	0.6724	+0.1886
» 13 . .	46.79	44.41	−0.32	−0.95	0.1024	0.9025	+0.3040
» 20 . .	46.65	44.26	−0.46	−1.10	0.2116	1.2100	+0.5060
» 27 . .	46.58	44.25	−0.53	−1.11	0.2809	1.2321	+0.5883
	M = 47.11	M = 45.36			1.1895	10.3272	Σ 3.4498

$$r = \frac{\Sigma(xy)}{n\sigma_x\sigma_y} = \frac{3.4498}{3.5046} = +\,0.984.$$

327. L'exemple qui suit est emprunté au matériel que nous avons réuni et travaillé dans notre étude sur « Les progrès économiques de la Belgique, de 1880 à 1908 » parue en 1911 dans le Journal de la Société royale de statistique de Londres. Nous avons comparé entre eux les indices de la production et les indices des échanges; à l'aide du même matériel nous avons déjà déterminé le résultat donné par la comparaison des courbes à l'aide de la formule $\frac{c-d}{c+d}$ (54). (Cfr. n° 313.) Maintenant nous allons procéder à un calcul plus complet et plus précis en recherchant le coefficient de covariation pour la série entière. La première colonne du tableau est réservée à l'inscription des années couvertes par l'observation. Les indices de la production et des échanges figurent aux colonnes 2 et 5; on remarquera que ces données sont des valeurs proportionnelles empruntées à un système d'Index-numbers; le coefficient de covariation se calcule de la même manière pour les données proportionnelles et pour les nombres absolus. La moyenne des indices est ensuite calculée : pour la production, nous la trouvons égale à $\frac{4278}{29} = 147.51$, chiffre que nous arrondissons à 148 pour la facilité du calcul; les indices des échanges ont pour moyenne $\frac{3788}{29} = 130.62$ que nous écrivons 131, afin de réduire le travail arithmétique. Les colonnes 3 et 6 reçoivent les valeurs des écarts, qui sont portées au carré aux colonnes 4 et 7 et multipliées entre elles à la colonne 8.

Ces éléments étant ainsi calculés et disposés, on procède comme précédemment et l'on trouve que le coefficient de covariation entre les indices de la production et des échanges est très élevé (0,979) (1) indiquant une union intime entre les deux phénomènes.

---

(1) Le professeur MORTARA, qui a calculé la même donnée, a obtenu le même résultat à un millième près, en moins.

EXEMPLE 6.

Indices de la production et des échanges en Belgique, de 1880 à 1908.

ANNÉES	PRODUCTION			ÉCHANGES			xy
	Index de la production	Variations par rapport à la moyenne (x)	x^2	Index des échanges	Variations par rapport à la moyenne (y)	y^2	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1880	102	— 46	2116	98	— 33	1089	+ 1518
1881	102	— 46	2116	98	— 33	1089	+ 1518
1882	110	— 38	1444	99	— 32	1024	+ 1216
1883	108	— 40	1600	101	— 30	900	+ 1200
1884	100	— 48	2304	100	— 31	961	+ 1488
1885	93	— 55	3025	99	— 32	1024	+ 1760
1886	92	— 56	3136	100	— 31	961	+ 1736
1887	100	— 48	2304	104	— 27	729	+ 1296
1888	105	— 43	1849	106	— 25	625	+ 1075
1889	115	— 33	1089	110	— 21	441	+ 693
1890	125	— 23	529	116	— 15	225	+ 345
1891	120	— 28	784	120	— 11	121	+ 308
1892	117	— 31	961	118	— 13	169	+ 403
1893	115	— 33	1089	119	— 12	144	+ 396
1894	120	— 28	784	122	— 9	81	+ 252
1895	122	— 26	676	125	— 6	36	+ 156
1896	141	— 7	49	126	— 5	25	+ 35
1897	150	+ 2	4	132	+ 1	1	+ 2
1898	163	+ 15	225	141	+ 10	100	+ 150
1899	186	+ 38	1444	144	+ 13	169	+ 494
1900	193	+ 45	2025	149	+ 18	324	+ 810
1901	163	+ 15	225	151	+ 20	400	+ 300
1902	179	+ 31	961	156	+ 25	625	+ 775
1903	198	+ 50	2500	162	+ 31	961	+ 1550
1904	211	+ 63	3969	166	+ 35	1225	+ 2205
1905	221	+ 73	5329	175	+ 44	1936	+ 3212
1906	247	+ 99	9801	183	+ 52	2704	+ 5148
1907	259	+ 111	12321	187	+ 56	3136	+ 6216
1908	221	+ 73	5329	181	+ 50	2500	+ 3650
	M = 148		69988	M = 131		23725	39907

$$r = \frac{\Sigma (x\,y)}{n\,\sigma_x\,\sigma_y} = \frac{39907}{40747{,}95} = 0.979$$

328. Le coefficient r, dans la forme où il est calculé plus haut est une expression purement numérique qui s'applique à la série considérée tout entière.

Le calcul repose tout d'abord sur la moyenne entre toutes les années qui composent la série; c'est d'après cette moyenne que les déviations sont calculées, multipliées entre elles, portées au carré, qui lui-même se trouve divisé par le nombre d'années d'observation, etc. Cette expression convient entièrement pour caractériser l'évolution respective de deux phénomènes considérés dans leur allure et dans leur direction générale, mais on peut concevoir que l'objet de la recherche porte sur un autre point. En effet, la direction générale des courbes n'est pas toujours le seul point à envisager : il arrive que pour obtenir une représentation exacte du phénomène il faille le décomposer et envisager séparément ce qu'on appelle les « mouvements séculaires », c'est-à-dire de très longue durée, et des variations plus ou moins rapides de grandeur, observables d'année en année et présentant fréquemment un caractère périodique. La méthode la plus facile pour observer ces variations courtes est due à M. Hooker; elle a été indiquée et appliquée dans un travail publié en 1901 portant sur la covariation entre la nuptialité et l'ampleur des transactions commerciales extérieures, en Angleterre (1).

Lorsque l'on compare les deux courbes de la nuptialité et du commerce extérieur par tête d'habitant (importations et exportations réunies), on remarque que les deux courbes ne présentent pas la même allure; tandis que celle du commerce n'a cessé d'aller en montant, celle de la nuptialité a subi un changement de direction depuis 1873 et sauf un relèvement passager entre 1895 et 1900 a cessé de garder une direction ascendante. Ce fait est très bien mis en évidence

(1) Hooker (R. H.), « Correlation of the Marriage — Rate with Trade ». (*Journal Royal Statistic Society*, 1901, p. 485.)

par le coefficient r qui, calculé de la manière ordinaire, indique 0.18 comme coefficient de covariation entre la courbe du commerce extérieur rapportée à la population et le taux de nuptialité. Ce coefficient très faible ne comporte qu'une seule interprétation : à savoir qu'il n'existe pas de rapport entre les deux phénomènes.

On aurait tort de s'en tenir à cette modalité de la recherche, car en appliquant une autre méthode, d'une grande simplicité, M. Hooker établit que la relation entre les deux phénomènes doit être exprimée par $r = 0.80$. Il n'y a pas de contradiction entre les deux résultats : en effet, le premier chiffre obtenu (0.18) marque qu'entre la direction générale des deux faits il n'y a pas de rapport, tandis que la seconde donnée signifie qu'il existe une relation étroite entre les oscillations des deux phénomènes. Or, c'est bien le second coefficient qui fournit la réponse la plus intéressante à la question que nous nous sommes posée : ce qu'il importe de savoir c'est si les échanges du commerce extérieur exercent une influence sur le taux des mariages pendant une période plus ou moins brève plutôt que d'être informé de la relation constante qui pourrait exister entre les deux faits. La méthode suivie par M. Hooker nous fournit les éléments de la réponse et c'est à ce point de vue qu'elle mérite de retenir l'attention.

Pour mettre en évidence la relation existante, M. Hooker a proposé d'établir les déviations non par rapport à la moyenne générale, mais sur des moyennes « instantanées » prises pour une période donnée. L'étendue du temps à considérer est déterminée d'une façon empirique en tenant compte de l'intervalle qui sépare les crêtes des oscillations, par exemple 5, 7, 9 ou 11 années. La moyenne des valeurs pendant cette période est calculée et le résultat est inscrit en regard de la valeur primitive centrale, c'est-à-dire que si la courbe montre une période de n années, la moyenne « instantanée » est portée en regard de l'année formant le centre de la période. La déviation est ensuite calculée entre

la donnée nouvelle et la première valeur, on procède ensuite comme pour le calcul ordinaire de la covariation en prenant soin de calculer la moyenne des σ_1 et σ_2 en tenant compte du nombre d'années pour lesquelles les valeurs nouvelles ont été obtenues et en négligeant celles ayant pour valeur 0.

329. Nous avons nous-même appliqué cette méthode à l'étude de la relation entre les indices de la production et des échanges en Belgique de 1880 à 1908, en utilisant les matériaux déjà mis en œuvre dans notre exemple 5. (Cfr. n° 326.)

Exemple 7. — **Indices de la production et des échanges en Belgique, de 1880 à 1908.**

Moyenne instantanée de cinq années (Méthode Hooker)

Années	PRODUCTION				ECHANGES				xy
	Index de la production	Moyennes instantanées ou mobiles	Variations par rapport aux moyennes (x)	x^2	Index des échanges	Moyennes instantanées ou mobiles	Variations par rapport aux moyennes (y)	y^2	
1880	102	—	—	—	98	—	—	—	—
1881	102	—			98	—	—	—	—
1882	110	104 4	— 5.6	31 36	99	99 2	— 0 2	0 04	+ 1.12
1883	108	102 6	- 5 4	29 16	101	99 4	+ 1 6	2 56	— 8.64
1884	100	100 6	- 0 6	0 36	100	99 8	+ 0.2	0 04	- 0.12
1885	93	98 6	— 5 6	31 36	99	100 8	- 1 8	3.24	+10 08
1886	92	98 0	— 6 0	36 00	100	101.8	— 1 8	3 24	+10 80
1887	100	101 0	- 1 0	1 00	104	103.8	+ 0 2	0 04	— 0 20
1888	105	107 4	2 4	5 76	106	107 2	— 1 2	1 44	+ 2 88
1889	115	113 0	+ 2 0	4 00	110	111 2	1 2	1 44	2 40
1890	125	116 4	+ 8 6	73 96	116	114 0	+ 2 0	4 00	+17 20
1891	120	118 4	+ 1 6	2 56	120	116 6	+ 3 4	11.54	+ 5 44
1892	117	119 4	- 2 4	5 76	118	119 0	— 1 0	1 00	+ 2 4
1893	115	118 8	3 8	14 44	119	121 6	— 2 6	6 76	+ 9 88
1894	120	123 0	- 3 0	9 00	122	122 0	0		0
1895	122	129 6	7 6	50 16	125	124 8	+ 0 2	0 04	- 1 52
1896	141	139 2	+ 1 8	3 24	126	129 2	- 3 2	10 24	— 5 76
1897	150	152 4	— 2 4	5 76	132	133.6	— 1 6	2 56	+ 3 84
1898	163	166 6	3.6	12 96	141	138 8	+ 2 2	4.84	— 7 02
1899	186	171 0	+15 0	225 00	144	143.4	+ 0 6	0 36	+ 9 00
1900	193	176 8	+16 2	262 44	149	148 2	+ 0 8	0 64	+12 06
1901	163	183 8	20 8	432 64	151	152 4	— 1 4	1.96	+28 72
1902	179	188 8	9 8	96 04	156	156 8	— 0 8	0.64	+ 7 84
1903	198	194 4	+ 3 6	12 96	162	162 0	0	—	0
1904	211	211 2	0 2	0 04	166	168 4	— 2 4	5 76	+ 0 48
1905	221	226 0	- 5 0	25 00	175	176 6	- 1 6	2 56	+ 8 00
1906	247	231 8	+15 2	231 04	183	178.4	+ 4.6	21 16	+60 92
1907	259	—	—		187	—	—	—	—
1908	221	—	—		181	—	—	—	—
				1602 00				85 50	165.00

$$r = \frac{\Sigma (xy)}{n\sigma_x \sigma_y} = \frac{165.00}{354.95} = + 0.464$$

Le résultat du calcul montre que les variations de la production et des échanges apparaissent unies par un lien beaucoup plus lâche quand on considère une période de brève durée, que quand on s'attache à déterminer la direction générale des faits; c'est la conclusion opposée à celle à laquelle M. Hooker avait été conduit, cependant la covariation conserve le signe positif.

Puisque le degré de covariation est élevé quand on envisage la période entière et qu'il devient plus faible quand on remplace la moyenne générale par une moyenne instantanée portant sur cinq ans, il va de soi qu'en adoptant une moyenne mobile comprenant un plus grand nombre d'années, onze par exemple, on doit obtenir un coefficient intermédiaire entre les deux valeurs obtenues précédemment. Nous avons procédé aux calculs nécessaires en utilisant une moyenne instantanée de onze ans, et nous sommes arrivé aux chiffres suivants :

$$\begin{aligned} \Sigma\,(xy) &= 484.62 \\ n &= 19 \\ \sigma_1 &= 13.206 \\ \sigma_2 &= 3\ 68 \\ r &= 0.5248 \end{aligned}$$

Le résultat obtenu confirme donc l'hypothèse émise plus haut.

330. Les méthodes précédentes sont applicables au cas dans lequel les variables consistent en une seule expression comprenant la classe tout entière; la plupart des séries chronologiques sont dans ce cas. Mais un autre procédé s'impose quand il s'agit de séries comprenant des variables partagées par classes de grandeur : âges, taux de salaire, taille, périmètre thoracique, etc., par exemple. Ces variables sont extrêmement nombreuses, la méthode pour en déterminer le degré de covariation ne manque pas de complexité et exige des calculs assez laborieux.

M. Yule en a donné une démonstration mathématique très claire que nous utilisons en partie ci-après avec la notation littérale adoptée par l'auteur.

Dans le calcul du coefficient r appliqué à deux variables comprenant plusieurs classes, il faut procéder comme dans la recherche de la moyenne par le procédé indirect (Cfr. n° 229), c'est-à-dire déterminer deux valeurs arbitrairement choisies et exprimer la grandeur des écarts par l'éloignement des classes de fréquence à partir de la valeur arbitraire choisie comme origine.

De même que dans le cas de la recherche de la moyenne, appelons ξ les déviations de l'une des variables à cette origine arbitraire, et η les déviations de la seconde variable. On appelle coordonnées des moyennes les valeurs des intervalles de classes à déterminer, multipliées l'une par l'autre avec leur signe : pour les distinguer des déviations, on les désigne par les lettres (surmontées du signe —) $\bar{\xi}$ pour la première variable, $\bar{\eta}$ pour la seconde.

On sait que la somme des déviations à partir de l'origine arbitraire est égale à la somme des déviations ($\Sigma\, xy$) par rapport à la moyenne, augmentée de l'intervalle de classe entre les deux valeurs multipliées par le nombre de termes ou d'intervalles. En conséquence :

$$\Sigma\,(\xi\eta) = \Sigma\,(xy) + \mathrm{N}\,\bar{\xi}\,\bar{\eta} \qquad (71)$$

ou :

$$\Sigma\,(xy) = \Sigma\,(\xi\eta) - \mathrm{N}\,\bar{\xi}\,\bar{\eta} \qquad (71)$$

De même on pose que si p représente les produits moyens calculés sur la moyenne ordinaire et p' les mêmes produits établis d'après l'origine arbitraire, on a :

$$p = p' - \bar{\xi}\,\eta \qquad (72)$$

Ces équations très simples servent de base à la méthode. Nous décrivons celle-ci avec quelque détail à raison de l'importance et de la complication qu'elle présente et pour plus

de clarté nous décomposerons les différentes parties du travail arithmétique appliqué à l'exemple choisi.

331. Cet exemple est celui de la relation existant entre l'âge du mari et l'âge de la femme au moment de la célébration du mariage. Les données originales, qui sont ici limitées entre les données — 21 ans et 45-50 ans à cause du manque d'homogénéité des classes, se trouvent exposées plus haut, n° 320. Nous adoptons comme origine arbitraire pour les femmes le centre de la colonne 25-30, soit 27,5 et pour les hommes, le centre de la ligne 30-35 = 32,5; le point de jonction de ces deux origines arbitraires tombe à la fréquence 2646.

Soit à rechercher la valeur de ξ par laquelle nous désignons l'intervalle de classe entre la moyenne de l'âge des femmes et la valeur arbitraire choisie pour cette catégorie. Dans ce but, nous disposons les fréquences (femmes) dans l'ordre des classes en adoptant la valeur 0 à l'origine arbitraire 27,5 (f = 11,967) et nous effectuons les produits. La somme algébrique des produits forme le numérateur d'une fraction dont le dénominateur est le nombre total de fréquences (42.841).

Nous avons donc :

$$\Sigma\,(f\bar{\xi}) = -\,29.479 + 14.003 = -\,15.476$$

valeur de la classe d'intervalle $= \dfrac{-\,15.476}{42.841} = -\,0.361$

Nous procédons de même pour $\bar{\eta}$ qui représente l'intervalle de classe entre la moyenne de l'âge des hommes et la valeur arbitraire choisie pour cette catégorie et nous obtenons :

$$\Sigma\,(f\bar{\eta}) = -\,45.056 + 9.208 = -\,35.848$$

valeur de la classe d'intervalle $= \dfrac{-\,35.848}{42.841} = -\,0.836$

d'où il résulte que $\Sigma\,(\bar{\xi}\bar{\eta})$ est de valeur positive et vaut :

$$-\,0.361 \times -\,0.836 = +\,0.302\ (0.301796)$$

Il faut ensuite déterminer les valeurs de σ_1 et de σ_2; on procède comme on l'a vu plus haut pour le calcul de la standard déviation.

Pour la première variable (femmes) le calcul nous donne :

$$\frac{\Sigma\,(f\,\bar{\xi}^2)}{N} = \frac{74.842}{42.841} = \sigma_1^{\,2} + 1.729$$

L'intervalle de classe $-\,0\,361^2 = \dfrac{+\,0.130}{+\,1.599}$

$$\sigma_1 = \sqrt{1.599} = +\,1.264$$

Les mêmes calculs sont établis pour σ_2^2 ; on a :

$$\frac{\Sigma\,(f\,\eta^2)}{N} = \frac{96.054}{42.841} = \sigma_2^2 = +\,2.242$$

$$-\,0.836^2 = \frac{+\,0.699}{+\,1.543}$$

$$\sigma_2 = \sqrt{1.543} = +\,1.242$$

332. Ces calculs ont déterminé la valeur des constantes $\bar{\xi}$, $\bar{\eta}$, σ_1, σ_2. Il reste à faire la recherche du coefficient *r* au moyen d'une table de corrélation dans laquelle sont indiquées, non seulement les fréquences, mais encore les déviations par rapport à l'origine arbitraire admise pour chacune des variables. On trouvera cette table à la suite de ces lignes, mais auparavant il faut faire remarquer que les valeurs positives sont groupées dans le quartier supérieur à gauche et le quartier inférieur à droite, tandis que les valeurs négatives sont placées dans le quartier supérieur à droite et dans le quartier inférieur à gauche. Cette disposition est déterminée par la combinaison des signes : en effet, les valeurs positives pour les femmes sont à droite, les valeurs négatives à gauche; pour les hommes, les valeurs négatives sont dans la partie supérieure du tableau et les valeurs positives dans la partie inférieure. La combinaison des signes fournit donc la disposition en croix que

nous avons signalée, comme le lecteur pourra s'en assurer par un simple schéma graphique (1).

EXEMPLE 8. — **Ages de la femme.**

Ages des hommes	-21	21-25	25-30	30-35	35-40	40-45	45-50	TOTAL
-21	774 VI	426 III	128 O	40 III	26 VI	2 IX	1 XII	1,397
21-25	3,401 IV	6,086 II	2,157 O	409 II	96 IV	39 VI	25 VIII	12,213
25-30	2,073 II	6,739 I	5,720 O	1,343 I	382 II	133 III	49 IV	16,439
30 35	498 O	1,669 O	**2,646 O**	1,601 O	442 O	180 O	78 O	7,114
35 40	140 II	441 I	838 O	852 I	512 II	221 III	105 IV	3,109
40-45	40 IV	170 II	344 O	380 II	291 IV	250 VI	133 VIII	1,608
45-50	21 VI	54 III	134 O	195 III	187 VI	208 IX	162 XII	961
TOTAUX.	6,947	15,585	11,967	4,820	1,936	1,033	553	42,841

(1) Le schéma suivant peut aider le lecteur à suivre le travail effectué au tableau de l'exemple 8.

F — H — } +	F + H — } —
F — H + } —	F + H + } +

333. Pour trouver le total des fréquences, on procède comme suit à l'aide du tableau ci-dessus en tenant compte de la valeur et du signe de $\xi\eta$: On inscrit dans un tableau spécial, à la première colonne les $\xi\eta$ qui, dans le cas actuel sont 1, 2, 3, 4, 6, 8, 9 et 12 et l'on fait la somme algébrique des produits positifs et négatifs trouvés dans chaque quartier en la multipliant avec son signe, par la valeur de $\xi\eta$; on fait ensuite la somme des fréquences positives, négatives et en croix, de manière à retrouver le total des fréquences. Les résultats du calcul sont indiqués ci-après :

Fréquences

$\xi\eta$	+	—	=	$(f\xi\eta$
I.	6739 + 852	1343 + 441	+ 5807	= + 5807
II	6086 + 2073 + 512 + 830	382 + 170 + 409 + 140	+ 7950	= + 15900
III.	426 + 221 + 195	133 + 40 + 54	+ 615	= + 1845
IV.	3401 + 291 + 105	96 + 49 + 40	+ 3612	= + 14448
VI.	774 + 250 + 187	26 + 39 + 21	+ 1125	= + 6750
VIII.	133	25	+ 108	= + 864
IX.	208	2	+ 206	= + 1854
XII.	162	1	+ 161	= + 1932
	22995	3411		+ 49400

On retrouve le total des fréquences :

$$22995 + 3411 + (11967 + 7114 - 2646) = 42841$$

La somme algébrique des fréquences par $\xi\eta = 49.400$.

$$p' = \frac{49\,400}{42.841} = 1.153$$

$$\bar{\xi}\bar{\eta} = +(-0.361 \times -0.836) = +0.302\ (0.301796)$$

$$p = p' - \bar{\xi}\bar{\eta} = 1.153 - 0\ 302 = +0.851$$

$$r = \frac{+0.851}{1.264 \times 1.242} = + \frac{0.851}{1.569} = 0.542$$

La covariation des âges des époux, dans les limites considérées (-21 à 45-50) est donc égale à + 0.542. Si elle était absolue, elle équivaudrait à l'unité ; dans le cas où elle serait nulle, elle serait exprimée par 0; si elle était en sens in-

verse elle serait précédée du signe — et pourrait prendre toutes les valeurs jusqu'à — 1. D'après la valeur de r (+ 0.542) il existe certainement une relation entre les âges des époux au moment de la célébration du mariage mais cette relation n'est pas très étroite et laisse place à de nombreuses exceptions.

Il est à remarquer que l'intervalle de classe doit être ici pris pour l'unité étant donné que l'unité est ici cinq ans ; il n'y a donc pas de réduction à effectuer.

334. Après avoir étudié les moyens de calcul mis en œuvre dans la recherche de la covariation de deux faits, le lecteur est à même de mieux se rendre compte de la valeur pratique du procédé. Le secours qu'il apporte au chercheur dans la détermination des relations entre phénomènes est indiscutable. On peut poser qu'en dehors des cas de parallélisme parfait, il est radicalement impossible d'apprécier par la vue seule avec exactitude, c'est-à-dire avec leurs relations mathématiques, les mouvements de deux courbes. Les statisticiens qui ne font pas usage des procédés décrits dans le présent chapitre, particulièrement du plus parfait d'entre eux, le calcul du coefficient de covariation, sont livrés à toutes les hésitations et condamnés à rester dans l'imprécision, à moins qu'ils ne soient conduits à conclure dans le sens de leurs opinions subjectives, ce qui est plus fâcheux encore. En voici un exemple typique : voulant déterminer la corrélation entre la tendance au mariage et l'état de prospérité économique, M. Hector Denis, après avoir rejeté les indices uniques tels que la variation du prix du blé et la consommation du charbon, écrit : « Je crois avoir trouvé une expression plus nette de l'influence de l'état économique sur la tendance au mariage. Elle est dans les variations de l'ensemble même des prix, dans les *index-numbers*. Avec la hausse générale des prix, l'esprit d'entreprise se développe, la demande de travail s'accroît, les salaires

s'élèvent, la tendance à contracter mariage devient plus vive, le nombre des naissances s'élève (1). »

Pour étudier la covariation des trois phénomènes, M. Hector Denis dispose un tableau à trois colonnes : l'index numbers de 28 articles exportés, sur la base 1866-1877 =100, le nombre de mariages par 1,000 habitants, le nombre de naissances par 1,000 habitants. Il est à remarquer que pour avoir une base de comparaison correcte, les nombres des mariages et des naissances auraient dû être calculés par le procédé des index-numbers. La même remarque est à faire à propos des graphiques. C'est sur ces éléments non comparables que M. Hector Denis a basé la conclusion que nous venons de lire.

335. Le calcul donne les résultats ci-après :

La covariation des index-numbers des prix et des mariages, envisagée pour la période entière 1850-1887 est exprimée par les données ci-après :

$$\Sigma\, xy = 1304.66$$
$$\sigma_x = \sqrt{105.67} = 10.279$$
$$\sigma_y = \sqrt{39.07} = 6.2507$$
$$n = 37$$

L'équation finale est

$$\frac{\Sigma\,(xy)}{n\,\sigma_x\,\sigma_y} = \frac{1304.66}{37 \times 10.279 \times 6.2507} = \frac{1304.66}{2377.28} = 0.548$$

Le coefficient de covariation est donc positif, mais on ne peut affirmer avec certitude qu'il existe une relation effective de cause à effet entre les deux phénomènes. Des statisticiens autorisés, comme M. G. Mortara, par exemple, n'admettent la signification du coefficient r que lorsqu'il

(1) HECTOR DENIS, *La dépression économique et sociale et l'histoire du prix.* Bruxelles, 1895, p. 150.

atteint une valeur numérique beaucoup plus considérable.

Il reste à rechercher si ce coefficient ne se modifie pas quand on considère une succession de mouvements de moindre étendue. Dans ce but, nous avons recours à la méthode de Hooker, exposée plus haut, et nous adoptons comme base du « trend » la moyenne prise sur 7 années consécutives. Nous avons dans ce cas :

$$\Sigma xy = 394$$
$$\sigma_x = \sqrt{25.843} = 5.083$$
$$\sigma_y = \sqrt{27.034} = 5.199$$
$$n = 29$$

Nous posons l'équation :

$$\frac{\Sigma (xy)}{n \sigma_x \sigma_y} = \frac{394}{29 \times 5.083 \times 5.199} = \frac{394}{766} = 0.514$$

Bien que le résultat auquel on arrive par l'emploi du « trend » soit un peu moins marqué que lorsqu'on envisage la relation générale entre les deux phénomènes, on peut affirmer, étant donnée la faible différence entre les résultats obtenus, qu'il y a identité de signification entre les deux méthodes employées.

L'induction tirée par M. Hector Denis de l'allure des séries paraît donc aboutir à une conclusion exagérée; il est nécessaire de recourir au calcul pour la corriger, la préciser et la ramener à ses proportions légitimes, qui laissent place au doute au lieu d'aboutir à formuler une sorte de loi générale.

V. — Equations de régression.

336. Sous le nom d'équation de régression on désigne un procédé d'investigation complémentaire à la méthode des corrélations.

Tandis que le coefficient de corrélation exprime numériquement le degré d'affinité qui peut exister entre la marche de deux phénomènes ou d'un plus grand nombre de faits

(Cfr. VI), les équations de régression permettent de déterminer les rapports de dépendance des mouvements des divers phénomènes entre eux. Elles répondent à la question suivante pour le cas de deux variables : à quel changement de valeur du phénomène A correspond le changement de valeur m du phénomène B ? S'il s'agit de trois variables, les équations de régression envisageront les changements de valeur de B et C par rapport à A, de A et C par rapport à B de A et B par rapport à C. Le coefficient de corrélation n'exprime donc que le degré d'affinité entre deux faits, tandis que les équations de régression font connaître quelque chose de plus, la répercussion que tout changement de valeur de A aura sur tout changement de valeur de B. Ainsi après les indications fournies par le coefficient de corrélation en ce qui concerne les rapports qui unissent deux phénomènes, les équations de régression viennent présenter un autre aspect de la même relation.

337. Il est désirable d'établir sommairement la genèse des équations de régression. Pour suivre cet exposé le lecteur voudra bien se reporter à la figure 28, qui représente par une ligne R R les moyennes des lignes horizontales dans une table de contingence et par une ligne CC les moyennes des colonnes verticales.

Soit M_2 la valeur moyenne de Y et soit RR coupant l'horizontale M_2x en M. Dans cette hypothèse, la verticale passant par le point M, coupera o X en M_1 (1). Or M_1 étant la moyenne de X, il s'en suit que la verticale passant par le point M_1 passe aussi par la moyenne de o X.

En effet, considérons l'inclinaison de RR par rapport à la verticale. La tangente (2) de l'angle M_1MR ou le rap-

(1) Cette théorie est basée sur le théorème suivant : D'un point pris hors d'une droite, on peut mener une perpendiculaire à cette droite, et on n'en peut mener qu'une.

(2) La tangente est la perpendiculaire élevée à l'extrémité du rayon qui passe par l'origine, jusqu'à la rencontre du rayon prolongé qui passe par l'extrémité de l'arc.

port de KL à LM est b_1. Admettons que les déviations observées à partir de MY et MX soient notées par x et par y. Il en résulte que pour chaque colonne horizontale du type y dans lequel le nombre des observations est n, nous aurons :

$$\Sigma\,(x) = n\,b_1\,y$$

et par conséquent, puisque $\Sigma\,(n\,y) = 0$ il viendra pour toute la table $\Sigma\,(x) = b_1 \Sigma\,(n\,y) = 0$. M, est donc la moyenne de X et M devient la moyenne de toute la distribution.

A l'aide du raisonnement qui précède, il est notamment établi que RR passe en M. Il reste maintenant à déterminer b_2. Cette détermination peut s'exprimer par les termes du produit moyen de tous les accouplements de lettres des déviations associés x et y.

D'où la formule :

$$p = \frac{1}{N}\,\Sigma\,(x\,y) \qquad (73)$$

Pour chacune des colonnes horizontales, il viendra :

$$\Sigma\,(x\,y) = y\,\Sigma\,(x) = nb_1\,y^2$$

et pour toute la table :

$$\Sigma\,(x\,y) = b_1\,\Sigma\,(n\,y^2) = N\,b_1\,\sigma^2_y$$

or

$$b_1 = \frac{p}{\sigma^2_y} \qquad (74)$$

De même si CC est la ligne sur laquelle tombent les moyennes des colonnes verticales et si b_2 est l'inclinaison par rapport à l'horizontale, r s/s M, il viendra :

$$b_2 = \frac{p}{\sigma^2_x} \qquad (75)$$

Il est d'usage d'inscrire les équations (74) et (75) sous la forme suivante :

$$r = \frac{p}{\sigma_x\,\sigma_y} \qquad (76)$$

En adoptant cette dernière notation, on en tire :

$$b_1 = r\,\frac{\sigma_x}{\sigma_y} \qquad b_2 = r\,\frac{\sigma_y}{\sigma_x} \qquad (77)$$

Dès lors, en écrivant les équations (49) en fonction de RR et de CC, nous aurons :

$$x = r\,\frac{\sigma_x}{\sigma_y}\,y \qquad y = r\,\frac{\sigma_y}{\sigma_x}\,x \qquad (78)$$

Ces dernières équations peuvent évidemment être exprimées dans les termes des variables X et Y plutôt que dans les termes des déviations x et y.

Quant aux équations (77) les quantités qu'elles expriment sont appelées les coefficients de régression ou, plus simplement les régressions, b_1 est la régression de x sur y ou l'écart en x qui correspond, en fonction de la moyenne, à un changement de une unité en Y. De même b_2 est la régression de y sur x, ou l'écart y qui correspond, en fonction de la moyenne, à un changement de une unité en x.

338. D'après les formules précédentes, nous pouvons calculer les équations de régression applicables aux données de l'exemple 6. (Cfr. n° 327.) Le coefficient entre les données concernant la production et celles relatives aux échanges est positif et très élevé : + 0·979·

Nous savons que

$$b_1 = r\,\frac{\sigma_x}{\sigma_y}$$

En remplaçant cette expression littérale par sa valeur numérique nous avons :

$$b_1 = +\,0.979\,\frac{49.126}{28.602} = +\,1.681$$

Nous avons vu que

$$b_2 = r\,\frac{\sigma_y}{\sigma_x}$$

D'où, en recourant aux données de l'exemple 6, nous tirons les valeurs numériques :

$$b_2 = +\,0.979\,\frac{28.602}{49.126} = +\,0.569$$

Comme l'a fait remarquer M. Udny Yule dans sa remarquable exposition de la corrélation entre le taux des salaires agricoles et la proportion de la population secourue dans trente-huit districts anglais (1) on peut, pour le calcul, remplacer la valeur des x et des y, c'est-à-dire les déviations à la moyenne, par les valeurs moyennes absolues des variables.

Recherchons d'après ce procédé, la valeur de l'équation b_1. Nous posons, d'après les calculs de l'exemple 6 cité plus haut :

$$X = 148$$
$$Y = 131$$

L'équation b_1 devient :

$$X - 148 = 1.681\,(Y - 131)$$

Simplifions et transformons l'équation. On a :

1° $X = +\,1.681\,(Y - 131) + 148$

2° $X = +\,1.681\,Y - [(1.681)\,(131)] + 148$

En effectuant la parenthèse, il vient :

3° $X = 1.681\,Y - 220.211 + 148$

4° $X = -\,72.211 + 1.684\,Y$

Cette équation est l'équation b_1 cherchée.

De même, nous pouvons substituer à y la valeur Y et rechercher dans ces conditions la valeur de l'équation b_2.

---

(1) Cfr. UDNY YULE, *loc. cit.*, p. 179.

Nous effectuons les calculs suivants :

$$Y - 131 = 0.569\ (X - 148)$$

y est ici remplacé par sa valeur Y 131.
x est remplacé par sa valeur X 148.
Nous simplifions et transformons :

1° $Y = 0.569\ (X - 148) + 131$
2° $Y = 0.569\ X + [(0\ 569)\ (-\ 148)] + 131$
3° $Y = 0\ 569\ X - 84\ 212 + 131$
4° $Y = -\ 46.788 + 0.569\ X$

Ce qui constitue l'équation b_2 cherchée.
Finalement nous obtenons les équations de régression :

$$X = -\ 72.211 + 1.681\ Y$$
$$Y = -\ 46.788 + 0.569\ X$$

Dans l'interprétation de ce résultat, il ne faut pas perdre de vue que les modifications se rapportent à un chiffre index 100 calculé sur l'année 1884.

VI. — Calcul des corrélations à trois variables.

339. Les calculs précédents concernent le cas où deux variables seulement se trouvent réunies. On peut concevoir la possibilité d'une telle rencontre lorsqu'il s'agit de phénomènes très simples, placés directement sous l'influence l'un de l'autre; mais lorsqu'il est question de faits plus compliqués, l'hypothèse de l'existence de deux variables seulement cesse d'être vraisemblable. La complexité des phénomènes économiques, particulièrement, atteint un degré élevé. On aimerait cependant à démêler les causes agissantes et à leur attribuer leur importance réelle. Le peut-on toujours à l'aide des calculs de la covariation? La réponse nous semble négative, tout d'abord parce que le nombre de causes reconnues est trop grand pour que, pratiquement, il soit possible de les soumettre au calcul; ensuite, parce qu'à mesure de l'augmentation du nombre des éléments de la formule, celle-ci perd de sa précision et

que son interprétation devient très délicate. Les résultats d'un calcul ne peuvent jamais être en contradiction avec la logique; on ne peut affirmer qu'il en serait toujours ainsi avec le calcul des covariations étendues à un nombre n de variables. La prudence dans les conclusions, qui a déjà été recommandée à propos de la covariation à deux variables est donc de mise, plus que jamais, quand il s'agit de trois variables et plus. Les statisticiens anglais qui ont créé la théorie de la corrélation à n variables le font expressément observer.

A moins de dépasser de beaucoup les limites que nous avons assignées au présent ouvrage dans le domaine des sciences mathématiques, nous ne pourrions aborder ici l'exposé de la théorie mathématique de la corrélation à n variables. Elle n'est cependant que la suite logique de la théorie de la corrélation à deux variables, mais elle est nécessairement beaucoup plus compliquée. Nous essaierons de donner une idée claire du travail arithmétique à effectuer pour la solution des équations, sans entrer dans l'exposé théorique de la méthode; le lecteur pourra de la sorte se faire une idée pratique de la question, mais pour la partie théorique et démonstrative, il devra nécessairement recourir aux traités de statistique mathématique et aux travaux spéciaux indiqués aux références : les ouvrages de MM. Pearson et Yule doivent être lus et relus par les travailleurs désireux d'approfondir ces matières.

340. L'exemple de covariation à trois variables que nous avons choisi comme illustration du travail arithmétique est tiré de la « Statistique des Industries extractives et métallurgiques » publiée par l'administration des mines en Belgique. Les chiffres de deux des données que nous allons utiliser ont déjà été donnés dans l'exemple 3 (Cfr. n° 316), où nous avons calculé le coefficient I de dépendance entre le bénéfice à la tonne et le salaire journalier moyen des ouvriers du fond entre 1885 et 1915. Aux deux données

dont il s'agit, nous en ajoutons une troisième : la valeur moyenne de la tonne de charbon. Il s'agit donc de déterminer quelle est la relation numérique existante entre ces trois éléments : le bénéfice à la tonne (X_1), la valeur moyenne à la tonne (X_2), le salaire journalier moyen de l'ouvrier du fond (X_3).

Dans sa forme la plus simple l'équation à l'aide de laquelle se calcule la corrélation à trois variables s'écrit comme suit :

$$r_{1.23} = \frac{r_{12} - r_{13}\, r_{23}}{\left(1 - r_{13}\right)^{1/2} \left(1 - r_{23}\right)^{1/2}} \qquad (79)$$

Les indices dont les coefficients de corrélation sont affectés sont l'objet d'une notation spéciale. Les premiers indices comprennent chacun deux chiffres sur trois chiffres existants concernant chacune des trois variables. Le premier des deux chiffres se rapporte à la variable indépendante, à la fonction; le second chiffre désigne la variable en fonction de laquelle les variations de la variable dépendante sont observées. Ainsi les coefficients de corrélation ci-après se rapportent :

r_{12}, au bénéfice à la tonne (X_1) en fonction de la valeur à la tonne (X_2)

r_{13}, au bénéfice à la tonne (X_1) en fonction du salaire journalier moyen (X_3)

r_{23}, à la valeur à la tonne (X_2) en fonction du salaire journalier moyen (X_3)

Les indices en question sont appelés *indices primaires;* on les appelle aussi *indices de l'ordre zéro.*

La suite du calcul a pour objet de déterminer les coefficients de corrélation du *premier ordre.* Ceux-ci sont affectés chacun de deux indices primaires et d'un indice secondaire; les indices secondaires désignent les variables restantes. Nous avons ainsi :

$r_{12.3}$ = corrélation de X_1, en fonction de X_2, reste X_3

$r_{13.2}$ = corrélation de X_1, en fonction de X_3, reste X_2

$r_{23.1}$ = corrélation de X_2, en fonction de X_3, reste X_1

341. Nous procéderons dans l'ordre des calculs et nous placerons sous les yeux du lecteur la suite des opérations.
En premier lieu, nous reproduisons les données numériques originales.

Bénéfice à la tonne, valeur à la tonne et salaire journalier moyen de l'ouvrier du fond dans les charbonnages de Belgique, de 1885 à 1911.

Années	Bénéfice à la tonne (X_1)	Valeur à la tonne (X_2)	Salaire journalier moyen (X_3)	Années
	Francs	Francs	Francs	
1885	0.40	8.87	3.20	1885
1886	0.30	8.25	2.94	1886
1887	0.48	8.04	2.99	1887
1888	0.65	8.43	3.10	1888
1889	1.11	9.45	3.42	1889
1890	2 84	13.14	4.20	1890
1891	1.82	12.58	4.18	1891
1892	0.60	10.28	3.56	1892
1893	0.33	9.35	3.29	1893
1894	0.40	9.32	3.25	1894
1895	0.40	9.45	3.43	1895
1896	0.52	9.50	3.49	1896
1897	0.90	10.26	3.72	1897
1898	1.05	11.00	3.94	1898
1899	1.71	12.43	4.37	1899
1900	4.26	17.41	5.21	1900
1901	2.33	15.23	4.69	1901
1902	1.41	13.20	4.39	1902
1903	1.23	12.99	4.38	1903
1904	0.75	12.59	4.20	1904
1905	0.83	12.64	4.28	1905
1906	1.91	15.00	4.98	1906
1907	2.16	16.86	5 52	1907
1908	1.37	16.14	5.17	1908
1909	0.74	14.37	4.64	1909
1910	0.50	14.59	4.85	1910
1911	0.14	14.76	4.96	1911

342. Il faut commencer par chercher les constantes de chacune des séries de variables. Par M nous désignons la moyenne de la série, par σ la standard-déviation, par r le coefficient de corrélation, chacune de ces lettres étant affectée de l'indice de la série de variables.

Nous calculons les M d'après la formule (3) :

$$M = \frac{1}{N} \Sigma (X)$$

et nous obtenons

$$M_1 = 1.153 \quad M_2 = 12.078 \quad M_3 = 4.087$$

Outre les moyennes, nous avons à calculer la standard-déviation de chacune des séries. Rappelons la formule (14A).

$$\sigma^2 = \frac{1}{N} \Sigma (X^2)$$

Nous reproduisons, page 545, pour guider le lecteur, les calculs servant à déterminer la standard-déviation de la série relative au bénéfice à la tonne (σ_1).

Nous calculons de même les standard-déviation de X_2 et X_3 et nous obtenons les constantes :

$$\sigma_2 = 2.768 \qquad \sigma_3 = 0.741$$

Il faut ensuite calculer les coefficients de corrélation existant entre :

1° X_1 et X_2; 2° X_1 et X_3; 3° X_2 et X_3.

Ces calculs, parfois assez laborieux, nous mettent en possession des constantes de corrélation qui s'inscrivent dans la première colonne du tableau ci-après (p. 548) intitulé : « Tableau du calcul de la corrélation à trois variables. »

343. Nous déterminons comme suit la première constante corrélative entre X_1 et X_2 et pour permettre au lecteur de suivre tous les détails de l'opération, nous reproduisons page 546 les calculs nécessaires pour déterminer r_{12}.

Calcul de σ_1

Années	Bénéfice à la tonne	Écarts (x) à la moyenne	x^2	Années
1885	0.40	0.753	0.567009	1885
1886	0.30	0.853	0.727609	1886
1887	0.48	0.673	0.452929	1887
1888	0.65	0.503	0 253009	1888
1889	1.11	0.043	0.001849	1889
1890	2.84	1.687	2.845969	1890
1891	1.82	0.667	0.444889	1891
1892	0.60	0.553	0.305809	1892
1893	0.33	0.823	0.677329	1893
1894	0.40	0.753	0.567009	1894
1895	0.40	0.753	0.567009	1895
1896	0.52	0.623	0.400689	1896
1897	0.90	0.253	0.064009	1897
1898	1.05	0.103	0.010609	1898
1899	1.71	0.557	0.310249	1899
1900	4.26	3.107	9 653449	1900
1901	2.33	1.177	1.385329	1901
1902	1.41	0 257	0 066049	1902
1903	1.23	0.077	0.005929	1903
1904	0.75	0.403	0.162409	1904
1905	0.83	0.323	0.104329	1905
1906	1.91	0.757	0.573049	1906
1907	2.16	1.007	1.014049	1907
1908	1.37	0.217	0.047089	1908
1909	0.74	0.413	0.170569	1909
1910	0.50	0.653	0.426409	1910
1911	0.14	1.013	1.026169	1911

$$\Sigma x^2 = 22.828803$$

$$\sigma^2 = \frac{22.828803}{27} = 0.845511 \text{ ou } 0.8455$$

$$\sigma = \sqrt{0.8455} = 0.919510739$$

Calcul de r_{12} (1 = bénéfice à la tonne, 2 = valeur à la tonne)

Années	BÉNÉFICE A LA TONNE			VALEUR A LA TONNE			xy	Années
	Importance du bénéfice	Ecarts par rapport à la moyenne (x)	x^2	Valeur à la tonne	Ecarts à la moyenne y)	y^2		
	fr.			r.				
1885	0.40	−0.753	0.567009	8.87	—3.208	10.291264	+ 2.415624	1885
1886	0.30	—0.853	0.727609	8.25	—3 828	14.653584	+ 3.265284	1886
1887	0.48	—0.673	0.452924	8.04	−4 038	16.305444	+ 2.717574	1887
1888	0.65	−0 503	0.253009	8.43	—3.648	13 307904	+ 1.834944	1888
1889	1.11	—0.043	0.001849	9 45	—2.628	6.906384	+ 0.113004	1889
1890	2.84	+1.687	2.845969	13.14	+1.062	1.127844	+ 1.791594	1890
1891	1.82	+0.667	0.444889	12.58	+0.502	0.252004	+ 0.334834	1891
1892	0 60	− 0.553	0.305809	10.28	—1.798	3 232804	+ 0.994294	1892
1893	0.33	−0.823	0.677329	9.35	—2.728	7.441984	+ 2.245144	1893
1894	0.40	− 0.753	0.567009	9 32	—2.758	7.606564	+ 2 076774	1894
1895	0.40	—0.753	0.567009	9.45	—2.628	6.906384	+ 1.978884	1895
1896	0.52	—0 633	0.400689	9.50	—2.578	6.646084	+ 1 631874	1896
1897	0.90	—0.253	0 064009	10 26	—1.818	3.305124	+ 0.459954	1897
1898	1.05	—0.103	0.010609	11 00	—1.078	1.162084	+ 0.111034	1898
1899	1.71	+0.557	0.310249	12.43	+0.352	0.123904	+ 0.196064	1899
1900	4.26	+3.107	9.653449	17.41	+5.332	28.430224	+16.566524	1900
1901	2.33	+1.177	1.385329	15.23	+3.152	9 935104	+ 3.709904	1901
1902	1.41	+0.257	0 066049	13.20	+1.122	1.258884	+ 0.288354	1902
1903	1.23	+0.077	0.005929	12.99	+0.912	0.831744	+ 0.070224	1903
1904	0.75	—0 403	0.162409	12.59	+0.512	0.262144	— 0.206336	1904
1905	0.83	—0 323	0.104329	12.64	+0.562	0.315844	— 0.181526	1905
1906	1.91	+0.757	0.573049	15.00	+2.922	8.538084	+ 2 211954	1906
1907	2.16	+1.007	1.014049	16.86	+4.782	22.867524	+ 4.815474	1907
1908	1.37	+0.217	0.047089	16.14	+4.062	16 499844	+ 0.881454	1908
1909	0.74	—0.413	0.170569	14.37	+2.292	5.253264	— 0.946596	1909
1910	0.50	—0.653	0.426409	14.59	+2.512	6.310144	— 1.640336	1910
1911	0.14	—1.013	1.026169	14.76	+2.682	7 193124	— 2.716866	1911

$\Sigma (xy) = 44.948884$

$n = 27$

$\sigma_x = 0.919$

$\sigma_y = 2.768$

$$r_{12} = \frac{\Sigma (xy)}{n.\sigma_x.\sigma_y} = \frac{44.948884}{68.682384} = 0.6544454$$

344. Calculés de la même façon, les coefficients de corrélation de X_1 avec X_3 et de X_2 avec X_3 ont respectivement pour expression :

$$r_{13} = 0.74992 \qquad r_{23} = 0,98556$$

La formule (68) a pour objet de permettre le calcul des corrélations de premier ordre et de leurs valeurs. Pour y arriver, il faut une suite assez longue d'opérations, dont le chercheur disposera, pour plus de facilité, les résultats sous forme de tableaux.

La détermination de $1—r^2$ donnerait lieu à des calculs très longs; heureusement il a été possible de construire des tables pour le calcul de la valeur de $1—r^2$; on en trouvera une très complète dans le précieux recueil de Pearson : « Tables for statisticians and biometricians » (1).

Nous cherchons ensuite les logarithmes de $(1—r)\frac{1}{2}$ au moyen de la table de Pearson; en employant une table de logarithmes à 7 décimales, nous arrivons à obtenir pour chacun des r dans l'ordre de leur énumération :

$$\bar{1},880776 \qquad \bar{1},82048905 \qquad \bar{1},7988476$$

Pour connaître le numérateur de la fraction, on fait d'abord la somme des produits en multipliant entre elles les valeurs des r restantes. Ainsi, nous avons pour résultat :

0,7350 0,6370 0,4875

valeurs qu'il faut soustraire des expressions corrélatives de l'ordre zéro, de sorte qu'on a :

— 0,1150 + 0,1130 + 0,4925

Nous déterminons immédiatement les logarithmes des numérateurs :

$$\bar{3},0606978 \qquad \bar{3},0565237 \qquad \bar{3},6924062$$

(1) Table VIII. Valeurs de $1 — r^2$ pour $r = 0,001$ à $0,999$ (pp. 20-21).

Les logarithmes des dénominateurs doivent être recherchés ensuite : le logarithme pour r_{12} s'obtient en additionnant les logarithmes de r_{13} et de r_{23} et de même pour les autres coefficients. On a donc :

$\bar{1},61933665 \qquad \bar{1},6796236 \qquad \bar{1},70126505$

Pour avoir les logarithmes de corrélation de premier ordre, on soustrait les dénominateurs des numérateurs et l'on obtient :

$\bar{1},4413612 \qquad \bar{1},3769001 \qquad 1,9911397$

Les valeurs de ces logarithmes sont :

0,28 0,48 0,98

On calcule en dernier lieu les logarithmes de $\sqrt{1-r^2}$ correspondant aux coefficients de corrélation de la colonne 8. A cet effet, nous utilisons comme la première fois, les tables de Pearson :

$$\log\ 0,921600 = 3,9645425; \quad \log.\ 0,769600 = 3,8862651;$$
$$\log.\ 0,039600 = 2,59769519$$

Tableau du calcul de la corrélation à trois variables

Constantes $\begin{cases} M_1 = 1,153 \ ; \ M_2 = 12,078; \ M_3 = 4,087 \\ \sigma_1 = 0,9196; \ \sigma_2 = 2,768; \ \sigma_3 = 0,741 \end{cases}$

r	log. $1-r^2$	Produits	Numérateur	Numérat. log	Dénomin. log.	Correlation de 1er ordre		log. (1/2) $1-r^2$
						log.	valeur	
1	2	3	4	5	6	7	8	9
$r_{12} = 0,65$	$\bar{1},880776$	0.7350	— 0 1150	$\bar{3},0606978$	$\bar{1}.61933665$	$\bar{1}.4413612$	0.28	$\bar{1}.98227125$
$r_{13} = 0,75$	$\bar{1},82048905$	0.6370	+ 0 1130	$\bar{3},0565237$	$\bar{1}$ 6796236	$\bar{1}.3769001$	0 48	$\bar{1}.94313255$
$r_{23} = 0,98$	$\bar{1},7988476$	0 4875	+ 0.4925	$\bar{3}.6924062$	$\bar{1}.70126505$	$\bar{1}.9911397$	0 98	$\bar{1}.29884759$

Ces résultats numériques obtenus, il reste à en donner l'interprétation.

345. Dans l'exemple proposé, nous désignons par :
(1) Le bénéfice moyen à la tonne;
(2) La valeur moyenne à la tonne;
(3) Le salaire journalier moyen de l'ouvrier du fond.

Nous essaierons de donner une interprétation rationnelle des coefficients de corrélation de l'ordre zéro et du premier ordre. Nous avons vu que :

$$r_{12} = 0.65 \qquad r_{13} = 0.75 \qquad r_{23} = 0.98$$

Le bénéfice moyen et la valeur moyenne à la tonne (1), tout en accusant une corrélation marquée, sont inférieurs au degré de corrélation montré par les autres éléments. Ce résultat est conforme à l'observation, car lorsque les prix sont élevés, on en profite généralement pour faire des travaux de premier établissement, dont le coût réduit l'importance du bénéfice.

La relation entre le bénéfice moyen à la tonne et le sa-

---

(1) La valeur à la tonne est le quotient de la division de la valeur globale du charbon extrait par le nombre de tonnes extraites. La valeur à la tonne ne représente donc pas le prix de vente moyen, car une partie assez importante de la production est consommée par les charbonnages eux-mêmes ou cédée à leurs ouvriers. Ainsi, en 1911, la production totale a été de 23,053,540 tonnes, dont la valeur atteignit 340,278,800 francs, ce qui établit la valeur moyenne à la tonne à fr. 14.76. D'autre part, il y a lieu de défalquer de ces chiffres 2,263,670 tonnes consommées par les charbonnages, d'une valeur de 20,082,100 francs (fr. 8.87 la tonne) ce qui laisse apparaître la quantité vendable à 20,789,870 tonnes et sa valeur à 320,196,700 francs ou fr. 15.40 la tonne. Ce qu'on appelle valeur moyenne à la tonne indique par conséquent le résultat brut de l'extraction.

Le salaire journalier moyen de l'ouvrier du fond est la somme nette payée à l'ouvrier travaillant au fond de la mine, défalcation faite des retenues prélevées sur les salaires pour des institutions de prévoyance, des amendes, fournitures d'outils, etc., divisée par le nombre de journées de présence de la masse des ouvriers du fond. Il en résulte que le salaire journalier moyen n'est pas un salaire effectivement gagné par des ouvriers réels. D'une part, il réunit tous les bassins, il confond toutes les exploitations en une seule; de l'autre, il réunit les salaires de ceux qui n'ont travaillé que d'une manière irrégulière aux salaires des ouvriers qui ont travaillé chaque jour et même ont fait des heures supplémentaires. C'est une pure abstraction dont la valeur de comparaison est seule à retenir.

laire moyen est plus fortement marquée : 0.75. L'indice de Fechner (indice de dépendance) donne un résultat très approché: 0.724. (Cfr. p. 497.)

La relation simple entre le salaire et le prix de vente apparaît très étroite, au point qu'elle se rapproche de l'unité : 0.98.

Les coefficients de corrélation du premier ordre sont les suivants :

$r_{12.3} = 0.28$ $r_{13.2} = 0.48$ $r_{23.1} = 0.98$

La signification de ces coefficients de corrélation est que, pour 12.3, le bénéfice moyen et la valeur moyenne à la tonne étant fonction l'une de l'autre, le salaire journalier moyen étant une variable indépendante, leur corrélation est la plus faible qui existe. Ainsi donc, si l'on considère uniquement ces deux éléments : la valeur à la tonne et le bénéfice réalisé, leur union apparaît assez étroite, mais si l'on y fait intervenir le salaire, la désunion apparaît aussitôt, car *l'élévation du bénéfice à la tonne, plus les autres frais, est en raison inverse de la hauteur du salaire* dans le mode de comptabilité qui a été adopté par l'administration des mines; or, la part du salaire est de loin la plus importante et c'est elle surtout qui pèse sur le reliquat constituant le bénéfice.

$r_{13.2} = 0.48$ indique que le bénéfice moyen à la tonne est en relation assez étroite avec le taux du salaire journalier,

---

Aux salaires s'ajoutent les frais provenant de travaux de premier établissement et autres frais, en supposant que les immobilisations, de quelque importance qu'elles soient, sont amorties immédiatement.

Le total de ces dépenses et des salaires, déduit du produit de l'exploitation, constitue le bénéfice global, qui, lui-même, divisé par le nombre de tonnes extraites, donne le bénéfice moyen à la tonne. Ce bénéfice n'est pas le même que celui accusé par la comptabilité des sociétés charbonnières, à cause des amortissements. L'ensemble de ces données ne représente évidemment qu'une valeur de comptabilité; il est clair que d'autres méthodes devraient être appliquées pour dégager les éléments sociaux des problèmes considérés (Cfr., par exemple, les publications de l'Office du Travail de Belgique sur les salaires dans l'industrie des mines).

la valeur à la tonne étant considérée comme variable indépendante.

Mais la conjoncture la plus favorable se présente lorsque la valeur à la tonne et le taux du salaire sont fonction l'un de l'autre, le bénéfice moyen étant variable indépendante; en effet

$$r_{23.1} = 0.98$$

Cette corrélation à trois variables permet d'arriver à une interprétation intéressante du mode de variation des salaires dans l'industrie des mines.

Dans une page lumineuse, Em. Waxweiler a décrit le mécanisme de ce qu'il appelle « le conflit des évaluations dans le débat du salaire (1) ». Au fond, le débat qui s'agite entre l'entrepreneur et le salarié chaque fois qu'il devient nécessaire de fixer le taux du salaire se ramène à ceci : pour l'employeur, le temps du travail doit être productif, pour le salarié, il doit être lucratif.

Il est évident qu'avant tout, les parties doivent être d'accord sur un salaire de base suffisant pour que l'ouvrier consente à travailler. La somme que représente ce salaire est variable dans l'espace et dans le temps; elle varie également d'après la catégorie sociale et technique du salarié. La somme qui déclenche l'activité d'un terrassier n'est pas celle qui déterminera un mécanicien ou un typographe à travailler; dans une même industrie et pour le même métier, le salaire courant ne sera pas identique en Flandre et en Wallonie. Le salaire courant est à peu près celui de l'ouvrier travaillant à l'heure ou à la journée. Mais pour le salaire aux pièces une série d'évaluations successives viennent modifier le taux habituel. Au fond, le salaire aux pièces n'est qu'un salaire au temps déguisé.

---

(1) *Bulletin de la classe des Lettres et des Sciences morales et politiques de l'Académie royale de Belgique,* Bruxelles, Hayez, 1907, page 147 et suivantes.

L'entrepreneur a pris soin d'étudier ou de faire étudier combien de pièces un ouvrier moyen peut faire durant une journée normale et il a fixé le prix unitaire de chaque pièce en conséquence. Le salarié accepte et travaille, mais en conservant par devers lui une certaine réserve de son activité. Peu à peu, il s'aperçoit qu'en disposant ses outils de telle façon plutôt que de telle autre, en évitant des pertes de temps, en adoptant tel tour de main suggéré par l'habitude et l'entraînement, il pourrait augmenter son gain journalier. Il réalise ces réformes et son salaire hausse. Dans bien des cas, il arrive que son salaire hausse dans des proportions que l'entrepreneur n'avait pas prévues. Or l'entrepreneur avait simplement cherché une base d'appréciation; cette base devient trop large; il réduit le taux unitaire de la pièce, ce qui semble à l'ouvrier une criante injustice, car l'entrepreneur lui semble profiter injustement des efforts qu'il a faits pour augmenter sa production.

Cette analyse est ingénieuse, mais elle omet de mettre en lumière un élément d'une importance réelle : la valeur marchande du produit sorti des mains de l'ouvrier.

Dans le travail des mines, cette valeur est aisément connue. Il n'en est pas de même dans les usines et dans les manufactures. Celles-ci produisent des centaines d'articles différents : une mine n'en produit qu'un seul : la houille. La valeur de la tonne est aisément connue : la Bourse, les journaux, le public font connaître à l'ouvrier les variations du prix du combustible. Reprenons à présent l'hypothèse initiale de M. Waxweiler. Supposons fixée la somme initiale, d'après les habitudes de la région, qui déclenchera l'activité de l'ouvrier mineur travaillant au fond. En admettant que l'ouvrier puisse faire tel avancement par jour, le prix de l'avancement est fixé. Survient une hausse du prix du charbon due à des demandes plus considérables de l'industrie. L'ouvrier mineur est au courant de cette hausse. De plus, chaque jour, en traversant la « paire », il constate que les stocks de charbon diminuent et que la

production de la mine s'écoule au fur et à mesure qu'elle arrive au jour. Le raisonnement que l'entrepreneur se fait lorsque l'ouvrier aux pièces arrive, par l'habitude, à dépasser d'une façon sensible le gain prévu lors de la fixation du prix unitaire, l'ouvrier mineur le refait *en sens inverse* : la base admise lors de la solution du conflit des évaluations n'est pas assez large, elle devrait être revisée, le prix de l'avancement sera augmenté; et sans augmenter son rendement, l'ouvrier gagnera un salaire plus fort.

Logiquement, il semblerait que les prétentions de l'ouvrier dussent être limitées par les bénéfices réalisés par le patron.

Mais au moment où le conflit d'évaluation éclate, l'ouvrier ne connaît pas le bénéfice du patron. Et l'entrepreneur l'ignore lui-même; c'est seulement à la clôture du bilan, au mois d'avril ou de mai de l'année suivante, que le patron connaîtra le bénéfice laissé par l'exploitation. C'est la raison pour laquelle dans les mines, le salaire et la valeur à la tonne sont fonction l'un de l'autre, et que l'intervention du bénéfice comme troisième variable ne modifie pas la relation existant entre ces deux éléments : la corrélation de l'ordre zéro est en effet la même que celle du premier ordre : 0.98.

346. *Références.*

BRAVAIS (A.), « Analyse mathématique sur les probabilités des erreurs de situation d'un point ». *Acad. des Sciences : Mémoires présentés par divers savants*. IIe série, t. IX, 1846, p. 246.

GALTON (F.), « Correlations and their Measurement ». *Proc. Roy. Soc.*, vol. XLV, 1888, p. 135.

HERON (D.), « On the relation of Fertility in Man to social status ». *Drapers Co Research Memoirs*, London, 1906.

HOOKER (R. H.), « On the Correlation of the Marriage-rate with Trade ». *Journ. Roy. Stat. Soc.*, vol. LXIV, 1901, p. 485.

ID., « On the Correlation of Successive Observations illustrated by Corn-prices », *ibid.*, vol. LXVIII, 1905, p. 696.

ID., « The correlation of the Weather and the Crops, *ibid.*, vol. LXX, 1907, p. 1.

HOOKER (R. H.) et YULE (G. U.), « Note on Estimating the relative influence of two variables upon a Third ». *Journ. Roy. Stat. Soc.*, vol. LXIX, 1906, p. 197.

MARCH (L.), « Comparaison numérique de courbes statistiques ». *Journ. Soc. Stat.* de Paris, 1905, pp. 255 et 306.

NORTON (J. P.), Statistical studies in the New York Money Market; Mac Millan, New-York, 1902.

PEARSON (K.), « Regression, Heredity and Panmixia », *Phil. Trans. Roy.* Series A, vol. CLXXXVII, 1896, p. 253.

YULE (G. U.), « On the Theory of Correlation ». *Journ. Roy. Stast. Soc.*, vol. LX, 1897, p. 812.

ID., « An Introduction to the Theory of Statistics ». London, Griffin, 1911, ch. IX, X, XI et XII.

ID., « On the Correlation of total Pauperism with Proportion of Out-relief ». *Economic Journ.*, vol. V, 1895, p. 603 et vol. VI, 1896, p. 613.

CHAPITRE V.

Statistique graphique.

I. — Définition et base géométrique.

347. Jusqu'à présent, nous avons vu la statistique faire usage exclusivement des chiffres pour exposer les résultats qu'elle obtient, mais il existe une autre méthode qui a recours aux figures géométriques pour exposer, comparer ou rechercher les conclusions de l'investigation scientifique : c'est la *statistique graphique.*

Sans l'aide de la statistique, les phénomènes collectifs resteraient pour nous un complexus indéchiffrable, car leur infinie variété nous interdirait à jamais la connaissance de la norme et des déviations par rapport à la norme. La statistique graphique vient encore compléter les acquisitions scientifiques de la statistique numérique en en traduisant les résultats au moyen de figures géométriques : pour connaître la situation d'un groupe de salariés, il faut calculer la distribution des salaires ; c'est l'œuvre de la statistique numérique ; puis on peut tracer la courbe de cette distribution, c'est la tâche de la statistique graphique.

Les raisons de recourir aux graphiques sont nombreuses :

1° Les chiffres, pour beaucoup de personnes, sont rebutants à lire. Cet inconvénient est supprimé ou atténué d'une manière notable si on leur substitue une figure géométrique simple ou une courbe statistique. Les figures employées par la statistique graphique sont des réalités qui reposent l'esprit de l'effort d'abstraction qu'il doit faire

pour saisir et comprendre les nombres. L'impression qu'elles produisent est en général plus vive et plus durable ;

2° Un plus grand nombre encore de personnes éprouvent une difficulté spéciale à retenir les chiffres ; ce n'est que par un effort dont bien peu sont capables qu'on parvient à apprécier avec exactitude les relations qu'ils présentent entre eux. Les courbes statistiques remplissent à cet égard un office d'une incontestable utilité ;

3° Les longues colonnes de chiffres sont remplacées dans les graphiques par une figure dont l'allure générale se découvre d'un seul coup. Cette méthode de la science possède donc un pouvoir de représentation synthétique qui n'appartient pas aux nombres ;

4° Les graphiques sont aussi un instrument de contrôle et de comparaison. Bien des inexactitudes passeraient inaperçues dans les séries numériques qui se décèlent du premier coup d'œil dans une courbe. Dans un grand nombre de cas on a eu recours à la statistique graphique comme à un moyen de signaler automatiquement les erreurs du relevé. Les comparaisons entre deux ou plusieurs faits observés sont aussi rendues plus faciles par l'emploi des graphiques ;

5° Les procédés graphiques ont aussi un rôle dans la recherche scientifique en faisant naître l'hypothèse nécessaire à la découverte de la loi : « En comparant les courbes de plusieurs faits portés sur un même diagramme, dit M. Levasseur, le statisticien, l'économiste, le moraliste découvrent souvent dans la similitude ou dans l'opposition des mouvements certains rapports qui leur avaient échappé, d'autres dont ils auraient eu peine, sans ce secours, à apprécier l'intensité ou la périodicité (1) » ;

6° Enfin, on peut recourir aux graphiques pour calculer

---

(1) Levasseur, « La Statistique graphique ». (*Jubilee volume of the Statistical Society*, London, 1885, p. 248.)

certaines données au lieu d'employer les méthodes numériques.

En résumé, la statistique graphique recherche, démontre et contrôle les résultats des phénomènes soumis aux investigations de la statistique. Nous trouvons dans ce triple rôle une indication au sujet de la division de notre matière : nous étudierons successivement les graphiques sous ce triple aspect en indiquant les méthodes applicables dans chaque cas.

348. La statistique graphique a un rôle à jouer qui lui appartient en propre ; elle n'est pas la servante de la statistique numérique, mais elle ne peut prétendre d'autre part à la précéder ou à la dépasser. De toute façon, la statistique doit commencer par recueillir les unités, les compter, les grouper, les peser ; c'est elle qui fournit aux constructeurs de graphiques la matière première dont ils ont besoin ; la statistique numérique peut, à la rigueur, se passer de la statistique graphique, mais celle-ci ne peut rien sans le concours de la première.

Dans un avenir plus ou moins prochain, les documents statistiques publiés par les gouvernements seront devenus tellement nombreux que les chercheurs éprouveront de grandes difficultés à se les procurer tous et à les utiliser. La statistique rétrospective deviendra très ardue et cependant son emploi sera plus indispensable que jamais. Les graphiques offriraient un moyen commode de sortir de ces embarras : il suffirait de traduire en courbes historiques, les principaux faits se rapportant aux années antérieures et de faire précéder de ces graphiques les statistiques nouvelles. Ainsi le présent serait relié au passé. Cet usage des graphiques aurait l'avantage de supprimer les recherches et de désencombrer les bibliothèques, double bienfait que les chercheurs de l'avenir, accablés sous le fardeau d'une littérature scientifique énorme, apprécieront de plus en plus.

349. Nous avons dit plus haut que les graphiques sont basés sur la géométrie. Bien que les figures employées soient en général très simples, il ne sera pas inutile de reprendre ici, sous une forme abrégée, l'exposé des principes mathématiques sur lesquels se fonde la statistique graphique; leur rappel est nécessaire pour permettre au lecteur d'envisager tous les cas théoriques et pour procéder avec sûreté dans les applications diverses. Les traités d'algèbre et de géométrie analytique contiennent tous l'exposé de la théorie dont nous ne faisons ici que rappeler les traits essentiels, le lecteur se reportera aux traités eux-mêmes pour avoir une idée plus complète des principes s'il désire approfondir la matière.

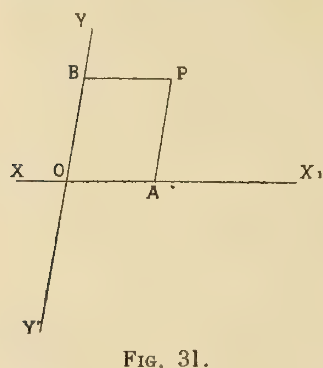


Fig. 31.

Pour fixer la position d'un point P dans un plan, il suffit de rapporter la position de P à des éléments supposés fixes dans le plan même. Dans ce but, on trace d'abord deux droites XX' et YY' se coupant dans ce plan; d'ordinaire, en statistique graphique, les deux droites en question se coupent à angle droit tandis que dans les figures employées en géométrie analytique, elles se coupent à un angle inférieur à 90°, mais il n'y a là aucune différence essentielle; celle qui vient d'être signalée n'est que de convention et d'usage. Ces droites sont nommées axe des ordonnées, pour la droite YY' placée dans le sens vertical, axe des abscisses pour la ligne XX' dont la position est horizontale. Le point de rencontre des deux axes est désigné par la lettre O et est pris par convention comme origine des distances. Par le point P, on mène des parallèles aux droites XX' et YY', les parallèles à ces droites s'arrêtent au point A sur l'axe des abscisses, au point B sur l'axe des ordonnées. La distance OB est appelée ordonnée du point P et la distance égale PA

porte le même nom ; la distance O A est l'abcisse de P ; les longueurs O A et P A considérées simultanément sont les coordonnées du point P. Il suit de là que la position d'un point P dans un plan est connue si l'on donne ses coordonnées.

350. Les données provenant du relevé sont d'habitude toutes de nature positive ; cependant il arrive qu'on recueille des données négatives, inférieures à un niveau donné, comme les indications du thermomètre, qui sont positives ou négatives d'après leur position par rapport à 0° ; le calcul arrive souvent à déterminer des données négatives, comme c'est le cas pour le produit ($x\ y$) dans les corrélations entre données groupées. Il importe donc d'envisager la position des valeurs positives et négatives dans le plan. On regarde comme positives les ordonnées mesurées au-dessus de l'axe XX' et comme négatives celles qui sont situées au-dessous ; on considère comme positives les abcisses comptées à droite de l'axe YY' et comme négatives celles qui sont portées à gauche.

Faisons une application de ces conventions à des données positives de part et d'autre. Soit un point P dont les coordonnées sont égales à $x = 2$, $y = 3$. Une unité étant adoptée pour la représentation des valeurs numériques, on reporte sur l'axe O X deux fois cette valeur et on la compte trois fois sur l'axe O Y ; les parallèles tirées de ces points $x\ y$ se coupent en un endroit qui est la position du point P. Les données étant toutes positives, il suffit de tracer les axes seulement jusqu'à leur origine com-

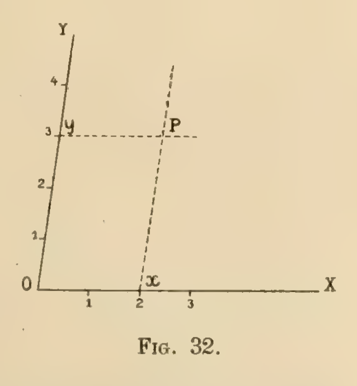


Fig. 32.

mune *o;* cette disposition est celle qui, d'ordinaire, est admise pour les diagrammes orthogonaux.

Supposons le cas de valeurs positives et négatives réunies et ensuite de valeurs négatives accouplées.

Soit les coordonnées $x = 3$, $y = -3$, puis les coordonnées $x = -2$, $y = -3$. La figure à tracer comporte le prolongement des axes X et Y au delà du point d'origine *o*. La longueur correspondant à l'unité étant admise, — on porte sur l'axe X à droite de l'origine *o* (puisqu'il s'agit de valeurs positives) une longueur égale à trois fois l'unité; puis sur l'axe Y, en dessous de *o*, une longueur 3; menant ensuite les parallèles à OX_1, OY_1 on les voit se couper en un point Q qui correspond aux coordonnées $x = 3$, $y = -3$.

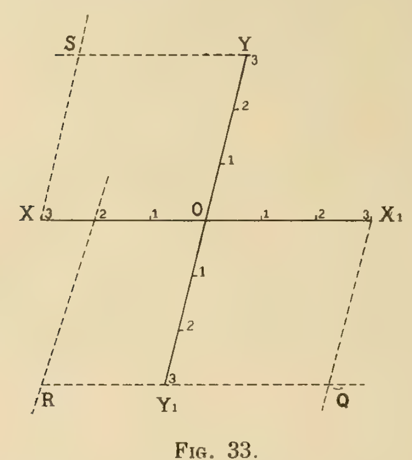


Fig. 33.

Pour trouver le point correspondant à $x = -2$, $y = -3$, on porte sur XX_1, à gauche de l'origine *o*, une longueur 2 et sur YY_1, en dessous de XX_1 une longueur 3; les parallèles menées comme il est dit ci-dessus, se coupent à l'endroit du point R déterminé par les coordonnées $x = -2$, $y = -3$.

Supposons le cas où les coordonnées sont égales à $y = 3$, $x = -3$, nous avons à mesurer une longueur 3 sur les X, à gauche de l'origine et sur l'axe des ordonnées Y une distance 3, à gauche de cet axe. Menant les parallèles comme précédemment, nous trouvons le point S correspondant aux coordonnées $y = 3$, $x = -3$.

D'où il suit que :

Pour deux valeurs positives de X et Y, la figure se trouve à droite au-dessus de l'axe des X.

Pour deux valeurs négatives de X et Y, la figure se trouve à gauche, au-dessous de l'axe des X ; pour une valeur positive et une valeur négative si X est positif et Y négatif la figure se trouve à droite et au dessous de l'axe des X ; elle se trouve à gauche et au-dessus de l'axe X si X est négatif et Y positif. Les principes fixés ci-dessus règlent la construction des diagrammes orthogonaux.

351. On peut aussi avoir à traduire graphiquement une fonction algébrique ou trigonométrique. Commençons par exposer le mode de représentation graphique d'une équation du premier degré, c'est-à-dire l'équation d'une droite. Nous examinerons d'abord le cas d'une équation du premier degré à une inconnue.

On sait par l'algèbre qu'on peut ramener une équation du premier degré à une inconnue à la forme

$$ax = b \text{ ou } x = \frac{b}{a}$$

et en représentant $\frac{b}{a}$ par m, nous avons :

$$x = m$$

laquelle équation représente une droite dont tous les points ont pour abcisse m; la ligne formée par ces points successifs est donc une droite parallèle à l'axe de Y. Si m est positif, on porte à droite de l'origine o une longueur égale à la valeur absolue de m estimée d'après une unité convenue et par ce point on mène une droite parallèle à O Y ; à gauche de l'axe YY_1 les valeurs de m sont négatives. Pour représenter une droite parallèle à l'axe des XX_1, on emploierait l'équation $y = + b$. Lorsque les équations sont : $y = 0$, $x = 0$, elles représentent l'axe même des X et des Y.

352. L'équation de la droite quand il s'agit d'une équation du premier degré à deux inconnues est un peu plus

compliquée. Soit a, b, c des quantités connues et $x, y,$ deux inconnues ; nous avons l'équation de la forme :

$$ax + by = c \quad \text{ou} \quad y = \frac{a}{b}x + \frac{c}{b}$$

que l'on transforme en

$$y = mx + n$$

Il convient d'abord de rechercher le lieu géométrique représenté par l'équation. Pour le trouver, si n est nul, il suffit de déterminer un point quelconque et de tracer la droite qui passe par ce point et l'origine. On démontre en effet que tous les points A, A', A'', A''' appartiennent à une même droite M N qui passe par l'origine, car :

Si X = O B et que nous prenons une longueur

$$Y = BA$$

on a

$$\frac{AB}{OB} = m$$

et A est un point du lieu géométrique cherché.

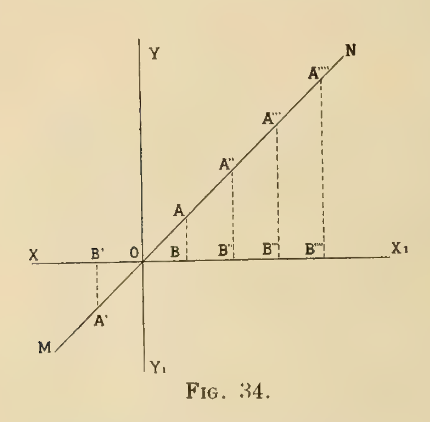


Fig. 34.

Si l'on donne à X une valeur négative

$$X = -OB'$$

Prenant Y = — B' A', on a :

$$\frac{-A'B'}{-OB'} = m$$

et A' est aussi un point du lieu.

On procède de même pour A'', A''', etc., et l'on obtient l'équation :

$$m = \frac{AB}{OB} = \frac{A'B'}{OB'} = \frac{A''B''}{OB''} = \frac{A'''B'''}{OB'''}$$

et tous les triangles formés par les points A, B, O, A', B', O, etc., sont égaux comme ayant un angle égal compris entre des côtés proportionnels.

Si n n'est pas nul, le lieu géométrique sera obtenu en augmentant ou en diminuant, suivant le signe de n, toutes les ordonnées de la valeur absolue de n.

Pour trouver les points où la droite coupe les axes, on fait $X = 0$, la valeur correspondante de Y donne le point où la droite rencontre l'axe des X ; si $Y = 0$, la valeur correspondante de X indique le point de section de Y.

Soit :

$$4x + by = 12$$

$$x = 0, \qquad y = 2; \qquad y = 0, \qquad x = 3$$

par un point 2 placé sur l'axe des Y on fait passer une droite qui sectionne l'axe des X au point 3. Cette droite AB est la droite exprimant l'équation ci-dessus.

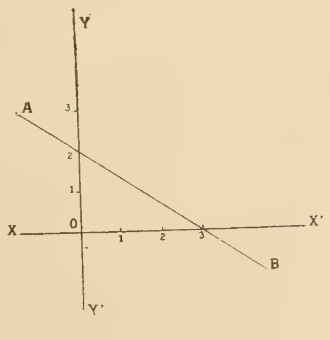


Fig. 35.

353. Les représentations graphiques du trinôme du second degré sont d'un usage très fréquent en statistique. Avant d'en indiquer la technique, il est utile de rappeler les bases mathématiques sur lesquelles elle repose.

Soit y la valeur du trinôme du second degré

$$x^2 - 4x + 3$$

Nous posons la fonction :

$$y = x^2 - 4x + 3$$

Faisons prendre à la variable x différentes valeurs. Si, en premier lieu, nous écrivons $x = 0$, il est évident que $y = + 3$. Il s'agit d'une valeur positive à chercher sur l'axe

des Y en se servant d'une longueur convenue représentative de l'unité; à la division 3, mesurée d'après cette unité, nous marquons le point A.

Si nous admettons $x = + 1$, nous avons évidemment $y = 0$. A la distance admise pour l'unité, nous marquons le point B sur l'axe des X, à droite de l'axe des Y, puisqu'il s'agit d'une valeur positive des X; les coordonnées de B sont $X = + 1$, $y = 0$.

Faisons encore varier X et admettons $X = + 2$. Dans ce cas, nous avons $y = - 1$; par une perpendiculaire à l'axe des X ayant son origine au point 2 nous posons le point C à une distance correspondant à — 1 sur l'axe des Y; les coordonnées du point C sont : $x = 2$, $y = - 1$.

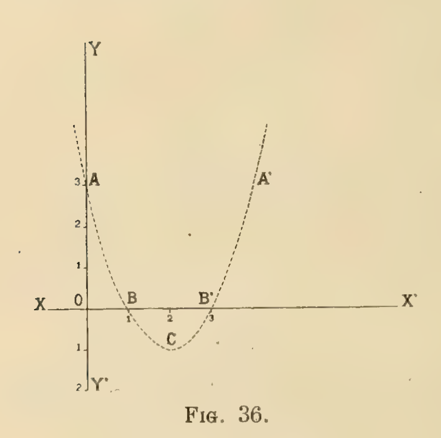


Fig. 36.

Supposons $x = + 3$, il en résulte $y = 0$, point représenté sur l'axe O X par B' (coordonnées $x = 3$, $y = 0$); pour $x = 4$, on obtient : $y = + 3$ (A'); pour $x = 6$ on a, $y = 15$, etc. Il suffit de relier par une ligne appropriée les points A, B, C, B', A' et l'on obtient ainsi une courbe dont les branches sont infinies.

La forme de la figure dépend de la nature des racines du trinôme : si les racines sont réelles, la figure conserve la forme indiquée plus haut et la courbe coupe l'axe des X en deux points marquant les racines de trinôme; si les racines sont égales, l'extrémité de la courbe est tangente à X X'; si elles sont imaginaires, la courbe ne touche pas l'axe des X X'. Ces formes géométriques se rencontrent en statistique.

354. Il nous reste à dire quelques mots de la fonction trigonométrique donnant naissance à la courbe sinusoïdale. Les faits statistiques exprimés par une courbe de cette nature sont contenus dans les séries périodiques : ce sont les séries qui montrent un phénomène observé selon les divisions du temps, dont les fluctuations maxima et minima se présentent à des intervalles réguliers. La représentation analytique de ces « vagues » oblige à avoir recours à des notions trigonométriques telles que le sinus. Le lecteur se rappelle qu'en trigonométrie on a établi la convention suivante : toutes les lignes trigonométriques des arcs terminés dans le premier quadrant sont positives, les lignes trigonométriques des arcs terminés dans les autres quadrants sont positives ou négatives suivant qu'elles ont les mêmes directions que les lignes de même nom des arcs du premier quadrant, ou des directions contraires. De quoi il résulte que le sinus est positif dans le premier et le deuxième quadrant et négatif dans les deux autres.

Des points déterminés d'après les règles qui précèdent, il résulte une figure périodique dont la courbe caractéristique porte le nom de *courbe sinusoïdale.*

355. On trouvera la table complète des sinus, degrés, minutes et secondes, dans les traités de trigonométrie. Pour tracer la courbe sinusoïdale, il suffit de disposer d'une table, contenant les sinus des 90 premiers degrés du cercle. Le lecteur sera à même de construire la courbe s'il dispose d'une table donnant simplement les sinus, par degrés. Nous la reproduisons ci-après :

Table des sinus de 0° à 90°

DEGRÉS	SINUS	DEGRÉS	SINUS	DEGRES	SINUS
0	0.000	30	0.500	60	0 867
1	0.017	31	0.515	61	0.876
2	0.035	32	0.530	62	0 884
3	0.052	33	0 545	63	0.892
4	0 070	34	0 559	64	0 900
5	0.087	35	0 574	65	0.908
6	0 105	36	0.588	66	0.915
7	0.122	37	0.602	67	0 922
8	0.139	38	0.616	68	0 928
9	0.156	39	0.629	69	0.935
10	0.174	40	0.643	70	0.941
11	0.191	41	0 656	71	0 946
12	0 208	42	0 669	72	0.952
13	0.225	43	0.682	73	0.957
14	0.242	44	0 695	74	0.962
15	0 259	45	0.707	75	0.967
16	0.276	46	0 721	76	0.971
17	0.292	47	0 733	77	0.975
18	0.309	48	0.745	78	0.979
19	0.326	49	0 757	79	0 982
20	0 342	50	0.768	80	0.985
21	0.358	51	0 779	81	0.988
22	0.375	52	0 790	82	0 991
23	0.391	53	0.800	83	0 993
24	0.407	54	0.811	84	0 995
25	0.423	55	0 821	85	0 996
26	0.438	56	0.831	86	0.998
27	0.454	57	0 840	87	0.999
28	0.469	58	0.850	88	0.999
29	0.485	59	0.859	89	1.000
				90	1.000

356. Pour construire la courbe sinusoïdale nous traçons un axe horizontal O X sur lequel nous portons à intervalles égaux les 360 divisions du cercle; en ordonnées, nous portons les divisions de l'unité : 0.1, 0.2, 0.3, etc., par chaque point de l'axe O X des abcisses, nous élevons des perpendiculaires d'une longueur égale à la mesure du sinus à ce point des abcisses, mesurée sur l'axe O Y des ordonnées. De ce qui précède, il résulte que les perpendiculaires sont placées au-dessus de l'axe O X (valeurs positives) pour les deux premiers quadrants et au-dessous de l'axe O X (valeurs négatives) pour les deux derniers quadrants. On obtient de la sorte la figure reproduite page 570, qui exprime théoriquement la courbe des phénomènes dont la périodicité serait idéale.

II. — La statistique graphique démonstrative.

357. On distingue dans la statistique graphique les diagrammes, les cartogrammes et les stéréogrammes.

Les diagrammes sont des figures construites d'après les procédés géométriques.

Les cartogrammes sont des cartes géographiques teintées ou ombrées selon l'intensité du phénomène à représenter.

Les stéréogrammes sont des représentations géométriques solides.

Les figures les plus usuelles sont les diagrammes; par leur variété, ils se prêtent à une quantité d'usages; leur précision les rend aptes à exprimer des relations mathématiques délicates et compliquées. Ils forment la partie la plus importante de la statistique graphique et en sont comme l'essence.

Les cartogrammes ne peuvent cependant être négligés. Leur emploi est requis chaque fois qu'il convient de lier à l'expression numérique d'un fait l'idée de sa localisation

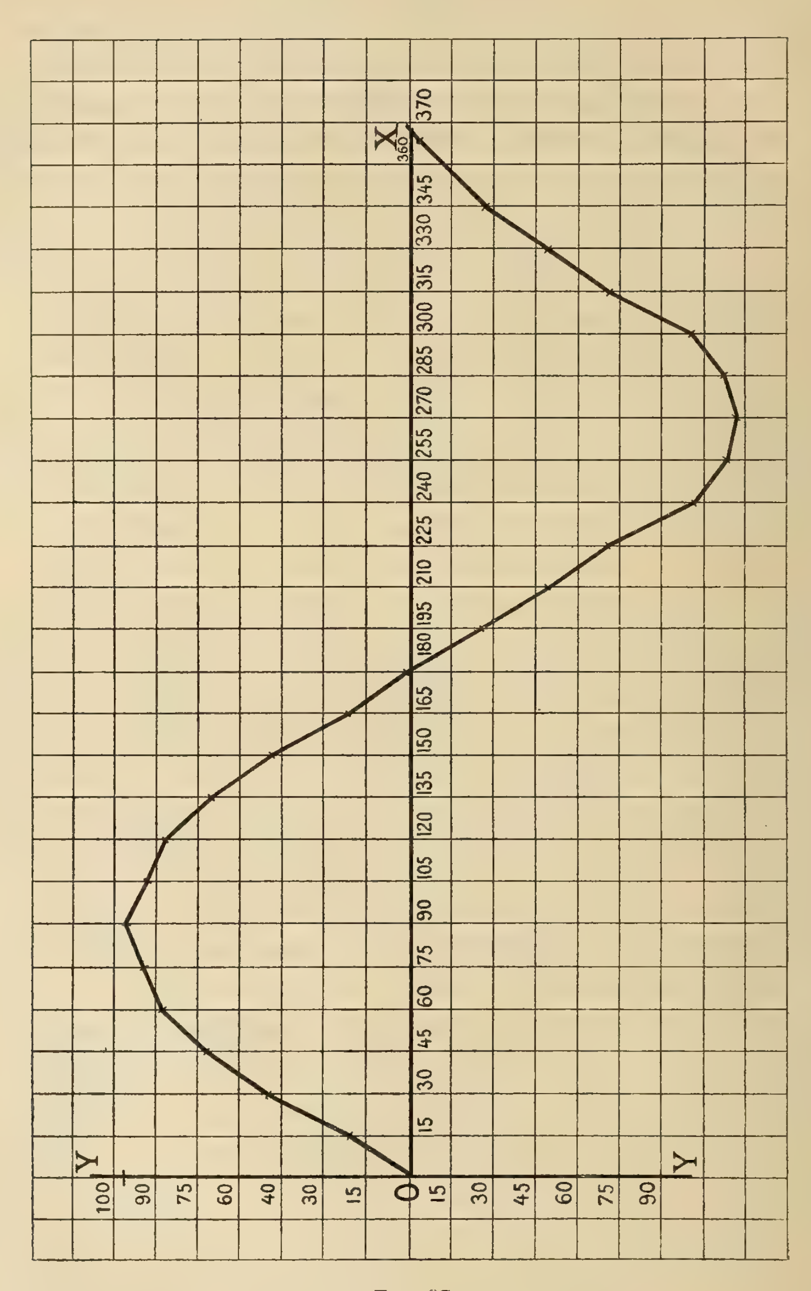

X
360
370
345
330
315
300
285
270
255
240
225
210
195
180
165
150
135
120
105
90
75
60
45
30
15
Y
Y
O
100
90
75
60
40
30
15
15
30
45
60
75
90

Fig. 37

géographique ; ils sont un moyen intuitif excellent d'apprendre la géographie économique.

Les stéréogrammes ou solides sont moins employés à raison de leur complication ; les spécialistes peuvent seuls y recourir avec fruit et ils se prêtent mal à l'enseignement.

A. — Diagrammes.

358. Les diagrammes comprennent un grand nombre de figures. Parmi elles, on distingue les figures représentatives simples (carré, rectangle, cercle) et les diagrammes orthogonaux dont le fondement géométrique a été exposé plus haut. Lorsque les diagrammes orthogonaux s'appliquent à des faits observés à certains intervalles de temps et que ces intervalles sont portés sur l'axe des abcisses tandis que la mesure de l intensité ou de la densité des variables est indiquée sur l'axe des ordonnées, ils portent le nom de diagrammes de succession, ou « historigrammes ». Quand les graphiques traduisent les données d'une table de contingence et que les variables sont portées, les unes sur les ordonnées, les autres sur les abcisses, ils sont appelés soit diagrammes ou polygones de fréquence, soit histogrammes ; ce dernier terme a été introduit en 1896 par le professeur Pearson (1). Les courbes dont il s'agit sont basées sur les nombres simples ; il en est d'autres qui dérivent des logarithmes de ces nombres et que pour cette raison on appelle « courbes logarithmiques ».

Enfin, certains diagrammes ont une base mathématique différente de celle des diagrammes orthogonaux, ils portent le nom de « diagrammes polaires ».

Les divisions introduites dans le présent exposé suivent les distinctions ci-dessus.

---

(1) K. Pearson, « Skew variations in homogeneous material ». Phil. Trans. Roy. Soc., 1896.

A. — *Figures représentatives simples.*

359. Les diagrammes représentatifs simples comprennent les procédés ci-après décrits :

1° *Le point.* Le mode figuratif par le point consiste essentiellement à disposer dans un certain ordre, des points ou petits cercles dont chacun représente une unité statistique ou un nombre convenu d'unités. En tant que procédé, ce mode de représentation ressemble au diagramme de surface et peut être avantageusement remplacé par celui-ci.

2° *La ligne.* On emploie soit une simple droite, soit des rectangles allongés dont la base est uniforme et dont la hauteur seule varie (tuyaux d'orgue). Pour ce diagramme, on commence par déterminer un rapport conventionnel entre l'unité métrique et une certaine quantité de faits à représenter, mais il n'est expressif que pour la comparaison avec une autre ligne ; ainsi, la longueur kilométrique du réseau de chemin de fer d'une série de pays, chaque réseau étant représenté par une ligne d'une longueur proportionnelle à l'étendue kilométrique. Veut-on exprimer un ensemble qui se compose de deux parties distinctes ? on emploie un trait plein pour une partie, un trait pointillé pour l'autre ; on peut aussi se servir de rectangles dont les parties sont diversement ombrées. Si l'on admet, par exemple, que la longueur du réseau ferré comprend les lignes d'intérêt local (chemins de fer vicinaux) aussi bien que les lignes d'intérêt général, on peut réserver à chaque fait une longueur proportionnelle à son importance. Les lignes exploitées par des compagnies et celles exploitées par l'Etat pourraient aussi être distinguées les unes des autres.

3° *Les surfaces.* Au lieu de représenter les faits par de simples droites ou des rectangles allongés, il est possible d'en figurer la grandeur relative au moyen de surfaces : deux éléments se trouvent ici à notre disposition, la lon-

gueur et la largeur, de sorte que la précision de la figure est augmentée en même temps que son pouvoir représentatif.

Commençons par éliminer de ce groupe une figure, le triangle isocèle, que certains statisticiens avaient cru pouvoir proposer pour exprimer les rapports de nombres sensiblement différents; la difficulté de se rendre compte de la valeur respective de surfaces ayant une base inégale a fait rejeter l'emploi de cette figure, en vertu du principe que la clarté de la représentation est la première condition d'un graphique bien établi.

Les figures de surface les plus usuelles sont le rectangle, le carré et le cercle.

On se sert des carrés et des rectangles soit pour comparer entre elles des données se rapportant à des faits différents, soit pour mesurer l'importance de faits ayant entre eux des rapports étroits. Dans ce dernier cas, les carrés ou les rectangles les plus petits sont inscrits dans le plus grand : dans l'autre cas, les figures sont séparées.

Les rectangles servent communément à indiquer deux dimensions, dont l'une est mesurée par leur base et l'autre par leur hauteur. Ils peuvent même servir à représenter trois valeurs si l'on admet que la troisième est fonction des deux premières : ainsi, on pourra figurer au moyen d'un rectangle le nombre de journées de travail le long du rectangle, le salaire quotidien à la base, la somme payée en salaires (aire). Le carré, dont les côtés sont égaux, est beaucoup moins expressif que le rectangle.

En général, il n'est pas aisé de comparer l'aire de deux carrés ou de deux rectangles dont les côtés varient; on éprouve la même difficulté à l'égard des aires des cercles. Aussi, réserve-t-on généralement le cercle pour figurer des données dont l'ensemble forme un total: par exemple la répartition de la population d'un pays par sexe, par groupes d'âges, par divisions professionnelles, etc. Chaque partie est d'abord comparée à l'ensemble et son importance rela-

tive est exprimée par rapport à cent : on prend ensuite sur la circonférence, une longueur proportionnelle (100=360°). Les différents secteurs sont ombrés ou coloriés de manière à les distinguer facilement les uns des autres.

360. Les diagrammes de longueur et de surface sont fréquemment employés dans les ouvrages de vulgarisation et, depuis quelques années, dans les livres scolaires, particulièment dans les manuels de géographie. On en a fait aussi un usage qu'on peut trouver surabondant dans les expositions générales et spéciales : des sections entières, comme celle d'économie sociale, ne sont formées que de graphiques, le plus souvent des diagrammes simples, dans la confection desquels les dessinateurs se sont ingéniés à mettre le plus de variété et de pittoresque possible, non sans détriment parfois pour l'exactitude de leurs figures. Il est permis de voir quelque exagération dans la mode qui s'est introduite de vouloir substituer aux chiffres des figures géométriques plus ou moins heureusement combinées. Le plus souvent, le diagramme, surtout celui basé sur les figures simples, reste inférieur à la donnée chiffrée parce qu'il lui est impossible d'atteindre la même précision : hormis des cas très rares, où il s'agit d'unités en nombre restreint, il n'est pas possible de faire varier un diagramme jusqu'à la dernière unité et si les unités sont peu nombreuses l'utilité du diagramme semble douteuse. Cette infériorité qui tient à la nature du diagramme se compense, il est vrai, par la vivacité de l'impression visuelle, supérieure chez beaucoup de personnes à l'excitation mentale produite par les chiffres; il convient de tenir compte de ce fait, tout en évitant l'abus du diagramme.

Ne serait-on pas d'avis qu'il y a abus du diagramme quand celui-ci vise à remplacer des données numériques simples qui, tout aussi aisément que les figures, peuvent se graver dans la mémoire? Serait-il plus difficile de retenir qu'en 1900-1901 la Grande-Bretagne comptait 41 millions

d'habitants, l'Autriche-Hongrie 45 millions et l'Empire allemand 56 millions que si l'on représente ces nombres par des carrés ou des rectangles d'une grandeur proportionnelle? Serait-ce alléger la tâche de la mémoire que d'obliger le lecteur à retenir que ces nombres peuvent être représentés graphiquement par des carrés ayant respectivement 6 cent. 4 mm., 6 cent. 7 mm. et 7 cent. 48 mm. de côté? De part et d'autre il s'agit de données numériques à confier à la mémoire et les secondes sont moins parlantes, moins suggestives que les premières.

Il y a abus encore quand le graphique s'éloigne des formes simples et mathématiques pour tomber dans les modes de représentation compliqués ou de pure fantaisie. L'essentielle qualité d'un graphique est d'être clair, ce qui exclut la complication voulue; son second mérite est d'être exact, ce qui écarte l'emploi des ornements d'imagination. Nous ne goûtons que médiocrement le procédé de ces géographes qui représentent le tonnage des flottes marchandes par une suite de vaisseaux de plus en plus grands, la puissance des armées par une ligne de soldats pygmées au début, géants à l'autre extrémité, la production de la laine ou du coton par des balles de marchandises de plus en plus grosses. Tout d'abord, nous n'apercevons pas la raison pour laquelle ces figures emblématiques se graveraient mieux dans la mémoire que les simples données numériques qu'elles expriment, d'ailleurs avec une précision très relative. Et, en second lieu leur exactitude est souvent sujette à caution : des vaisseaux, des hommes, des ballots sont des solides, or, il n'est pas certain du tout que le dessinateur en tienne toujours compte pour établir ses figures.

Bien que la recommandation puisse paraître superflue, nous notons ici que pour comparer entre elles des aires de surface, il faut établir les dimensions de ces aires dans le rapport des racines carrées des aires elles-mêmes s'il s'agit de surfaces, et dans le rapport des racines cubiques s'il s'agit de volumes.

B. — *Diagrammes orthogonaux.*

1° Diagrammes de succession ou historigrammes.

361. Lorsqu'un phénomène est étudié en relation avec le temps et que les divisions du temps forment les points de distribution, on a ce qu'on appelle une série historique : c'est l'histoire d'un fait résumé par dates et par chiffres. Au diagramme qui reproduit ces données en portant en ordonnées les nombres et en abcisses les divisions du temps, on a donné le nom d'historigramme. Les auteurs français n'ont pas adopté cette dénomination, qui vient de l'école anglaise; ils ont donné à ces diagrammes le nom de diagrammes de succession parce qu'ils marquent la façon dont un fait succède à un autre fait. Nous ne voyons pas de raison de donner la préférence à un terme plutôt qu'à l'autre, c'est pourquoi nous les emploierons chacun indifféremment.

Le diagramme de succession ou historigramme est le procédé le plus utile que la statistique graphique mette à notre disposition pour analyser les phénomènes complexes. Une longue série numérique défie pour ainsi dire l'analyse, à moins de recourir à des procédés compliqués que nous avons exposés plus haut. Au contraire, le tracé, même capricieux, d'un graphique forme un schéma que l'analyse atteint bien plus aisément que les chiffres.

La théorie des diagrammes orthogonaux est celle de la détermination d'un point dans un plan, que nous avons exposée plus haut, y compris la théorie de la courbe sinusoïdale qui s'applique aux phénomènes périodiques. Ces exposés bien que sommaires, sont suffisants pour envisager la question sous son aspect théorique; aussi nous plaçons-nous à présent au point de vue technique et pratique.

362. Pour mieux faire ressortir l'utilité spéciale des diagrammes à l'égard des séries historiques, il est désirable

de procéder à une application. Nous reprendrons les données d'un exemple déjà utilisé précédemment : le bénéfice moyen à la tonne réalisé par les charbonnages de Belgique de 1885 à 1915. Voici ces données :

EXEMPLE 1. — **Bénéfice moyen à la tonne, de 1885 à 1915, réalisé par les charbonnages en Belgique**
(*matériel extrait des rapports annuels de l'Administration des Mines*)

Années	Bénéfice à la tonne	Années	Bénéfice à la tonne	Années	Bénéfice à la tonne
1885	0 40	1895	0.40	1905	0.83
1886	0.30	1896	0.52	1906	1.91
1887	0.48	1897	0.90	1907	2.16
1888	0.65	1898	1.05	1908	1.37
1889	1.11	1899	1.71	1909	0.74
1890	2.84	1900	4.26	1910	0.50
1891	1.82	1901	2.33	1911	0.14
1892	0.60	1902	1.41	1912	0.34
1893	0.33	1903	1.23	1913	0.83
1894	0.40	1904	0.75	1914	0.63
				1915	0.75

Nous portons sur l'axe des abcisses, à des intervalles égaux, les divisions du temps, c'est-à-dire dans le cas actuel, les années. Les données mêmes de la série sont relevées d'après cette division et il n'y a place ici pour aucun choix. Il n'en est pas de même en ce qui concerne l'unité de valeur : le bénéfice moyen à la tonne varie dans des limites très étendues, puisque nous le voyons à fr. 4.26 en 1900 et à fr. 0.14 en 1911. La moyenne annuelle se fixe à fr. 1.806. Pour mieux observer la grandeur des oscillations nous adoptons pour unité des ordonnées la valeur de fr. 0.50; partant de zéro au point d'origine, nous portons

sur l'axe des ordonnées autant de divisions qu'il en faut pour atteindre la somme maximum de fr. 4.26. De chaque division de l'axe des abcisses nous situons un point dont la hauteur correspond au bénéfice de l'année d'après la graduation portée sur l'axe des ordonnées : nous complétons le diagramme en reliant tous ces points par des droites dont l'ensemble forme une ligne brisée.

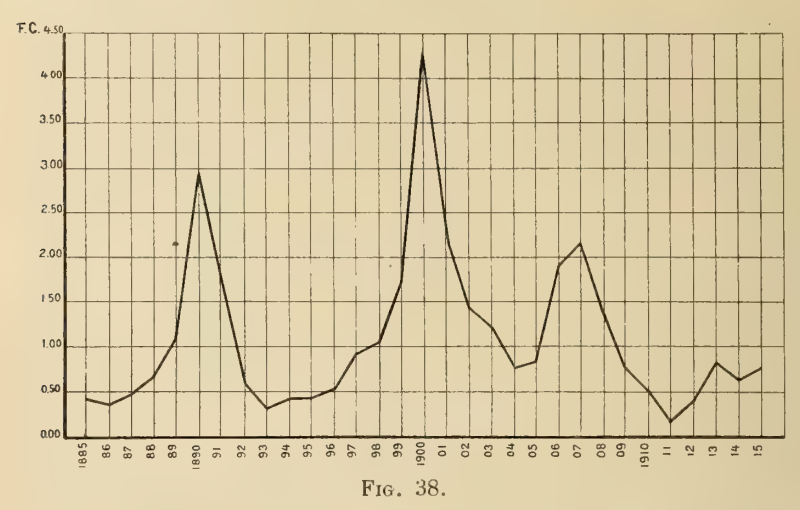


Fig. 38.

Le graphique fait apparaître les traits essentiels du phénomène beaucoup mieux et surtout beaucoup plus vite que ne le fait le tableau numérique. Au premier coup d'œil s'aperçoivent trois pics, séparés par des dépressions plus ou moins profondes : le pic compris entre les années 1889-1891, avec son sommet en 1890 ; un massif central plus important limité par les années 1898 et 1902 dont le sommet, très élevé, se place en 1900 ; une pyramide moins élevée que les sommets précédents comprise entre 1905 et 1908 avec son maximum en 1907. Les dépressions extrêmes se remarquent en 1886, 1893, 1904 et 1914. Les bénéfices des charbonnages en 1900 ont atteint à peu près cent millions de francs. La courbe correspond en gros aux années grasses et aux années maigres de la situation économique générale. Il y a cependant une restriction à apporter à cette

conclusion : elle provient de ce que, pour établir le bénéfice net des charbonnages, base de la redevance proportionnelle des mines, l'administration porte en compte chaque année toutes les dépenses, savoir, les dépenses extraordinaires aussi bien que les frais courants de l'exploitation. Le montant des installations nouvelles vient donc diminuer, d'une manière un peu factice, les bénéfices des charbonnages (1).

363. La construction des graphiques orthogonaux soulève différents problèmes techniques d'une solution parfois délicate.

Dans les diagrammes de succession ou historigrammes, les divisions du temps sont portées sur l'axe des abcisses. Ces divisions doivent être égales, mais rien n'est fixé quant à leur étendue. Tout ce qu'on peut dire à cet égard, c'est que la précision du diagramme est d'autant plus grande que la division du temps adoptée pour mesurer est petite. Supposez qu'un observateur dresse un relevé de la température heure par heure ; le diagramme dont l'unité temps serait l'heure serait beaucoup plus précis que celui dressé par vingt-quatre heures. Mais le statisticien n'est pas toujours libre d'adopter une échelle du temps aussi précise qu'il le voudrait. Ou bien les matériaux réunis par d'autres lui sont remis tout préparés et il n'est pas en son pouvoir de les modifier ; ou bien, s'il les réunit lui-même, la crainte de compliquer le relevé ou d'imposer au public une gêne trop grande fait donner la préférence à une unité de temps assez étendue.

Le résultat d'une investigation quelconque arrêtée à une date déterminée est ou une moyenne ou un total, ou une observation particulière faite à cette date même. Ce dernier résultat est le plus précis qu'on puisse obtenir, mais dans les sciences économiques et sociales il est rare que l'on puisse s'en contenter. Pour exprimer la continuité des phé-

(1) *Statistique générale de la Belgique.* « Exposé de la situation du Royaume, de 1876 à 1900 », vol. III, p. 220, Bruxelles, 1914.

nomènes sociaux, il faut nécessairement recourir aux moyens de représenter cette continuité : la moyenne qui est un partage, la totalisation qui est une accumulation. La droite tracée hypothétiquement d'une année à l'autre pour réunir les sommets des ordonnées est donc bien sujette à caution, et elle l'est d'autant plus que les intervalles sont espacés.

Si telle est l'exigence théorique, il ne faut pas oublier que l'avantage principal du graphique consiste dans sa puissance de concentration. A vouloir trop préciser on risquerait d'atténuer la valeur du schéma graphique. La faculté de résumer en une simple impression visuelle toute une situation n'appartient qu'à la statistique graphique. En jetant les yeux sur le diagramme de la page 578 on aperçoit du même coup les trois périodes d'augmentation des bénéfices, les époques de dépression, les années les plus favorables et les plus mauvaises, on sait juger de l'importance relative de ces trois périodes de hausse, de la grandeur de l'intervalle qui les sépare, de leur durée, etc. Pour dégager des tableaux de chiffres une impression aussi complexe, il faudrait beaucoup de temps ; le diagramme au contraire nous la fait apparaître en un instant.

364. Les considérations qui précèdent s'appliquent, dans leur ensemble, aux ordonnées comme aux abcisses, cependant l'échelle des ordonnées est en général plus précise, c'est-à-dire plus détaillée que l'échelle des abcisses. Les mesures quantitatives procèdent, d'ordinaire, par unités, alors qu'il est rare de voir adopter des divisions du temps plus courtes que l'année.

La question la plus délicate soulevée par la construction des diagrammes orthogonaux est le rapport à établir entre les intervalles des axes X et Y. Si les divisions de X (abcisses) sont largement espacées, alors que celles de Y (ordonnées), sont resserrées, la courbe s'étale paresseusement et ses « vagues » ressemblent aux flots tranquilles

d'un fleuve de plaine. Avec la convention contraire, les pics se hérissent, les sommets se dressent, les dépressions se creusent comme les lignes tourmentées de crêtes montagneuses à l'aspect chaotique. Les rapports mathématiques sont restés les mêmes, mais l'allure du diagramme est toute différente. Le premier donne l'impression d'une succession de phénomènes sans surprise ou à-coups, le second éveille l'idée d'une suite de faits extrêmement accidentée, presque de cataclysmes. L'Institut international de statistique, sur le rapport du Dr Jacques Bertillon, a recommandé d'adopter des échelles telles que l'allure moyenne du phénomène corresponde, pour la tangente de la courbe, à une inclinaison de 45°. En même temps il a signalé l'inconvénient qu'il y a à couper la partie inférieure du diagramme comme on le fait souvent sous prétexte qu'elle est inutile; cette suppression arbitraire fausse le diagramme en faisant croire que les variations de la fonction sont plus importantes qu'elles ne le sont réellement (1).

365. L'étude des conventions relatives aux rapports à établir entre les abcisses et les ordonnées a été reprise par l'Institut international de statistique dans une de ses sessions ultérieures. Comme le fait observer le rapporteur, M. L. March, tout jugement porté sur la variation relative des ordonnées dépend avant tout du rapport qui existe entre l'unité de mesure des abcisses et l'unité de mesure des ordonnées. Si les phénomènes représentés sont de même nature, il est facile de tracer toutes les courbes d'après la même échelle : elles seront alors exactement comparables entre elles. Il n'en est pas de même si l'on a à comparer des courbes se rapportant à des phénomènes de nature différente. Cette question des comparaisons sera étudiée plus loin. (Cfr. III : La statistique graphique comparative.) Pour

---

(1) *Institut international de statistique*, VIIIe session. « Propositions relatives à l'uniformité à apporter dans l'établissement des graphiques », par le Docteur Jacques Bertillon, p. 313.

l'instant, bornons-nous à ce qui a trait à la fixation de la relation conventionnelle entre la grandeur moyenne des ordonnées et l'unité de mesure des abcisses. L'Institut international de statistique, sur la proposition de M. March, a admis notamment de substituer aux nombres à représenter le rapport de toutes les ordonnées à leur valeur moyenne; afin d'éviter de recommencer les calculs de moyennes chaque fois qu'une année nouvelle vient s'ajouter aux autres, on est convenu de déterminer une moyenne de base portant toujours sur les mêmes années, par exemple 1901-1910. On recommande, dans ce cas, de représenter la grandeur de base par une longueur égale à celle qui représente trente années. Cette convention admise, dit M. March, la courbe étendue à partir de 0, sur cent années, aurait une pente de 3/10; limitée à 10 années seulement, elle aurait la pente de 3/1; pour 30 années, la courbe se trouverait inclinée à 45° et l'utilisation de la feuille de papier se trouverait la plus complète possible.

Dans un travail antérieur (1) nous avons appliqué cette méthode à l'illustration des données de la statistique commerciale, sauf que le point 100 sur l'axe des ordonnées représente cinquante années et non trente; ce dernier nombre convient mieux pour les phénomènes démographiques que pour les faits économiques. La moyenne annuelle des importations et exportations réunies de la Belgique de 1901 à 1910, a été de 5,637,950,833 francs. Cette moyenne est supposée égale à 100; une même longueur prise sur l'axe des ordonnées et sur l'axe des abcisses représente le point 100 et cinquante années; la longueur prise sur l'axe des ordonnées est partagée en 100 parties égales qui correspondent aux variations pour cent; tous les chiffres absolus, de 1831 à 1910, sont transformés ensuite en nombres proportionnels. Nous les donnons ci-après pour chaque année terminant une période décennale.

---

(1) « Le Commerce extérieur de la Belgique. » (« Etudes sur la Belgique », Anvers, 1912.)

ANNÉES	VALEUR ABSOLUE	PROPORTION %
1831.	186,543,841	3.21
1840.	345,239,643	6.12
1850.	431,955,770	7.66
1860.	986,944,911	17.51
1870.	1,610,901,760	28.57
1880.	2,897,633,275	51.39
1890.	3,109,139,044	55.14
1900.	4,138,637,146	75.40
1910.	7,672,389,012	136.09

L'allure du diagramme est aussi juste que possible, alors qu'il est si facile, avec la progression rapide des chiffres du commerce, de donner à la courbe une pente exagérée.

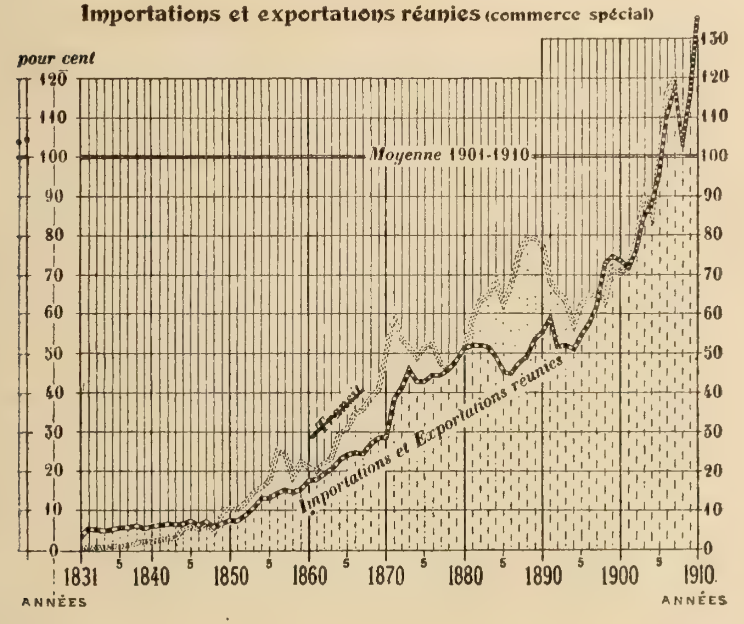


Fig. 39.

366. La matière présente assez d'importance au point de vue technique pour qu'il soit opportun de donner un second exemple accompagné d'un exposé détaillé et critique. Nous reprenons donc l'exemple que nous avons utilisé précédemment, relatif aux variations du bénéfice moyen à la tonne réalisé par les charbonnages belges de 1885 à 1915.

Nous commençons par calculer la moyenne annuelle des nombres compris dans la période 1901-10. Le calcul donne pour résultat 1.32; la valeur fr. 1.32 représentera donc le point 100 sur l'axe des ordonnées. La distance entre le point 100 et l'origine du système sera égale à la longueur portée sur l'axe des abcisses pour représenter la période des 31 années observées, si l'on adopte entièrement la convention proposée par l'Institut international de statistique. Si l'on préfère admettre que le point 100 représente 50 années, la longueur portée sur les abcisses sera de 31/50 de celle portée sur les ordonnées. Tous les chiffres sont rapportés à la valeur de base fr. 1.32. Voici les résultats de ce calcul :

ANNÉES	%	ANNÉES	%	ANNÉES	%
1885	30.3	1895	30.3	1905	63.3
1886	22 7	1896	39.3	1906	144.7
1887	36.4	1897	68.6	1907	164.0
1888	49.2	1898	79.5	1908	103.7
1889	84.0	1899	121.9	1909	56.1
1890	215 1	1900	322.7	1910	37.8
1891	137.8	1901	176.5	1911	10.6
1892	45.4	1902	106.8	1912	25.7
1893	25.0	1903	93.1	1913	63.3
1894	30.3	1904	56.8	1914	47.7
				1915	56.8

Le lecteur, en utilisant une feuille de papier quadrillé pourra très rapidement tracer le diagramme. Il ne pourra manquer de faire la remarque que la forme du graphique est plus élancée que dans le diagramme construit d'après les nombres absolus, le maximum (1900 = 322.7) étant trois fois 2/10 plus élevé que la moyenne et que la distance couverte sur l'axe des abcisses par la série des années.

367. Cette remarque nous amène à dire quelques mots de l'influence de la position de la moyenne sur l'allure de la courbe.

Dans le cas actuel, si la moyenne, au lieu d'être établie sur les années 1901-10, l'avait été sur les années 1900-09, donc avec une avance d'une année seulement, elle eût atteint le chiffre de fr. 1.70 au lieu de fr. 1.32 et le pourcentage le plus élevé eût été de 250 seulement au lieu de 322. L'allure de la courbe aurait été très différente de celle qu'elle présente quand la courbe est établie d'après les chiffres proportionnels reproduits ci-dessus. La convention de prendre uniformément la moyenne sur les années 1901-10 présente l'avantage de ne pas nécessiter de nouveaux calculs à chaque année qui s'ajoute à la série, mais elle a aussi l'inconvénient de ne pas tenir compte dans tous les cas du véritable emplacement de la moyenne et de dénaturer alors l'allure de la courbe. Le lecteur en sera convaincu en jetant les yeux sur les calculs suivants qui se rapportent au sommet observé en 1900 et en déterminant la hauteur d'après différents systèmes de prendre la moyenne :

Moyenne 1901-1910 = fr. 1.32 = 100; 1900 = fr. 4.26 = 322.7
Moyenne 1900-1909 = fr. 1.70 = 100; 1900 = fr. 4.26 = 250.6
Moyenne 1885 1915 = fr. 1.87 = 100; 1900 = fr. 4.26 = 227 8

La convention admise par l'Institut international de statistique s'adapte bien aux phénomènes qui ne sont pas susceptibles de modifications importantes ou rapides, comme les phénomènes démographiques. Il n'en va plus de même quand les variations sont profondes et précipitées, comme

c'est souvent le cas dans les faits économiques; le graphique prendrait une allure plus exacte s'il était établi en adoptant pour base la moyenne de la série entière. Il est vrai qu'à chaque année on se trouve obligé dans ce cas, de recommencer le calcul de moyenne, mais ce n'est là qu'un inconvénient négligeable en regard de l'avantage d'une courbe dont l'allure est absolument exacte.

2° Diagrammes de distribution ou histogrammes.

368. Les diagrammes de distribution sont ceux qui ont pour but de faciliter l'étude d'une série de faits considérés à un moment, indépendamment de l'ordre dans lequel ils ont pu se produire. Ainsi, la division du temps, qui est un élément essentiel dans les diagrammes de succession ou historigrammes est exclue des graphiques dont il s'agit ici. Cette différence entre les diagrammes dont il a été question est essentielle mais il en est d'autres encore qu'on peut ranger sous trois chefs :

1° Quand il trace un historigramme, le statisticien ne doit porter sur l'axe des abcisses qu'un nombre de divisions relativement peu élevé, il est rare qu'il dispose de séries comprenant plus de cinquante divisions du temps. Au contraire, les diagrammes de distribution s'appliquent à des séries comportant souvent un grand nombre de degrés de variabilité à inscrire sur l'axe des abcisses, tandis que cette nécessité ne se fait pas sentir au même point pour le diagramme de succession; dans ce dernier, une division de l'axe des abcisses est réservée à chaque division du temps, le plus souvent l'année; au contraire dans le diagramme de distribution, comme on ne peut songer à porter sur l'axe des abcisses des divisions en nombre tel qu'elles fassent apparaître séparément chaque degré de variabilité du phénomène, il faut alors établir des classes parmi les variables. Nous avons exposé plus haut les règles théoriques de la division par classes.

2° Les diagrammes de succession ont une forme irrégulière, tandis qu'on a constaté parmi les courbes de dispersion une régularité assez grande pour que le professeur Pearson ait pu en calculer six types différents. (Cfr. n° 206.)

La courbe de distribution symétrique parfaite, dont nous parlerons au chapitre suivant, est la formule typique du polygone de fréquence exprimé au moyen du diagramme de distribution ou histogramme. Elle a pour formule :

$$y = \frac{n}{\sigma\sqrt{2\pi}} \cdot \frac{}{e^{x^2}/2\sigma} \qquad (80)$$

Au moyen de cette formule on peut calculer la valeur de chaque ordonnée y à toute distance x mesurée le long de l'axe X X' à partir du mode. La lettre e représente la base du système népérien de logarithmes : 2.71828 ; σ = la déviation type, n = le nombre total de variables ; π = 3.14159...

3° Les diagrammes de succession représentent simplement l'importance d'un fait unique, en rapportant cette donnée à l'époque à laquelle il s'est produit. Les diagrammes de distribution envisagent deux faits en fonction l'un de l'autre : des ouvriers et le salaire qu'ils gagnent, par exemple.

369. Les diagrammes de distribution peuvent être établis de deux façons :

A. Sur l'axe des abcisses on porte, à intervalles égaux, les divisions des différentes classes dont se compose la série. Sur l'axe des ordonnées on indique les divisions par fréquence ; pour chaque classe, on note la hauteur correspondante à la fréquence de la classe et on réunit les points par une série de droites. Le point est placé au milieu de l'intervalle de la classe. Le diagramme établi d'après ces bases porte le nom de *polygone de fréquence.*

B. On peut aussi élever sur chaque intervalle des classes un rectangle ayant cet intervalle comme base et dont la hau-

teur est égale à la fréquence de la classe. Ce diagramme a reçu le nom d'histogramme (K. Pearson).

Nous avons donné plus haut, en étudiant la théorie de la distribution, de nombreux modèles de polygones de fréquence et d'histogrammes : nous prions le lecteur de s'y reporter.

370. Presque toujours, les termes extrêmes de la série continue sont en nombre très limité, ainsi qu'il résulte des caractères généraux de la courbe des erreurs. La question se pose de savoir si l'on peut considérer ces termes extrêmes comme des accidents et ne pas les comprendre dans le calcul. On ne doit pas se résoudre à cette dernière solution sans de bonnes raisons. Pour le cas où l'on devrait procéder à cette élimination, voici quelle est la marche du calcul :

Reprenons l'exemple du « *Pimpinella saxifraga L* » que nous avons donné au n° 219. Nous avons les données suivantes :

Nombre de rayons : 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15.

Nombre d'exemplaires : 1, 5, 9, 22, 38, 62, 61, 29, 14, 4, 4.

$$M = 10.253 \qquad \sigma = 1.786$$

$$K = \frac{2n - 1}{4n} = \frac{497}{996} = 0.49899$$

D'après la table des valeurs de « probabilité intégrale normale » insérée dans les « statistical methods » de C. B. Davenport (table IV) nous voyons qu'au nombre 0.49899 correspond la valeur 3.08. La limite de déviation peut alors être considérée comme égale à $3.08 \times 1.786\ (\sigma) = 5.50088$. La classe-limite $= 10.253 - 5.501 = 4.752$.

Aucune observation ne tombant dans une classe inférieure à la classe-limite, il n'y a lieu d'exclure du calcul aucune des données de l'expérience.

371. Les conventions admises en ce qui concerne les courbes chronologiques peuvent s'appliquer avec quelques

modifications aux courbes de fréquence. De même qu'on l'a vu plus haut, l'Institut international de statistique a recommandé, pour les courbes de fréquence, de rapporter chaque ordonnée à l'ordonnée moyenne, ou à la somme des ordonnées, c'est-à-dire à la surface de la courbe. En vue d'utiliser le mieux possible la surface d'une feuille de papier, supposée carrée, M. L. March a établi que le résultat le meilleur serait atteint si l'on donnait à une amplitude E, comprenant la moitié de la surface, une longueur égale, sur l'axe des abcisses, au dixième de celle qui mesurerait le total des observations et que, d'autre part, la colonne médiane représentant la moitié des observations aurait comme hauteur, environ cinq fois l'amplitude E. Ces suggestions sont utiles pour réaliser l'unité en matière de graphiques, mais elles sont toujours subordonnées aux indications résultant de la nature du cas particulier à résoudre.

C. — Courbes logarithmiques.

372. Il semble, après l'exposé des systèmes graphiques qui vient d'être fait, que la statistique ait à sa disposition un instrument à la fois fidèle et précis pour traduire d'une façon schématique les mouvements et les modifications des phénomènes démographiques, économiques et sociaux. Mais si les modes habituels de représentation graphique peuvent suffire, dans le plus grand nombre de cas, à ce qu'on attend d'eux, il est cependant des circonstances où leur apparente précision pourrait n'être qu'une illusion.

Les diagrammes orthogonaux historiques basés sur les nombres absolus des variables peuvent facilement donner naissance à des erreurs d'appréciation. Leur échelle des ordonnées est établie d'après des divisions conventionnelles, ou ce qui vaut mieux, elle est basée sur une moyenne. Si la valeur unitaire, portée en ordonnée, est élevée, les oscillations de faible importance absolue seront à peine perceptibles bien que leur importance relative, à l'égard des don-

nées les plus proches, soit tout aussi grande que celle des ordonnées se rapportant à des nombres élevés. Bowley donne un exemple bien choisi, faisant toucher du doigt cette difficulté : le commerce d'exportation de l'Angleterre, dit-il, qui était de 52 millions de livres en 1815, tomba à 42 millions l'année suivante : cette diminution de 20 p. c. est presque imperceptible dans un diagramme historique s'étendant à la période séculaire du XIX[e] siècle, tandis que l'attention se porte immédiatement sur la dépression qui se marque entre 1883 où le commerce d'exportation atteignait une valeur de 305 millions et 1886 où il ne représentait plus que 269 millions, ce qui fait 12 p. c. de diminution.

La raison en est qu'avec l'échelle arithmétique des diagrammes orthogonaux, la position d'un point sur les ordonnées varie dans la même mesure aussi bien dans le cas d'une augmentation en valeur d'un million de francs succédant à une valeur de 10 millions, que dans le cas d'une augmentation d'un million après avoir enregistré une valeur de 100 millions. De part et d'autre, il y a une même valeur absolue qui se marque par un égal accroissement mesuré sur l'échelle des ordonnées, mais la première hausse d'un million représente 10 p. c. de la valeur précédente et la seconde n'atteint que 1 p. c. Or, l'importance relative des changements survenus au cours de la période historique importe plus, en général, que leur grandeur relative. Veut-on, par exemple, mesurer l'intensité des mouvements représentés par le diagramme, ce sont les changements relatifs que l'on consultera bien plus que les nombres absolus. Pour représenter graphiquement les modifications relatives d'un phénomène, on a recours aux courbes logarithmiques dans lesquelles on substitue aux nombres absolus, sur l'échelle des ordonnées, leurs logarithmes.

373. La construction des courbes logarithmiques présente certaines particularités sur lesquelles il est utile d'attirer l'attention. Le lecteur trouvera dans n'importe quel

traité d'arithmétique les éléments de la théorie des logarithmes, qu'il n'est donc pas expédient de rapporter ici. Toutefois, il convient de faire remarquer ici que les courbes logarithmiques n'expriment pas les variations absolues des nombres dont elles dérivent, mais bien leurs variations relatives. Ce sont, dit Bowley, des diagrammes de rapports, non de quantités (1). Ce point qui ne devra pas être perdu de vue, précise l'usage qu'il est permis de faire de ces courbes et les circonstances dans lesquelles elles peuvent être employées.

Prenons, comme premier exemple, les données déjà utilisées au n° 362 relatives au bénéfice moyen à la tonne réalisé par les charbonnages de Belgique de 1885 à 1915. Nous inscrivons, en regard des nombres absolus, leurs logarithmes. Le lecteur se rappellera que les nombres inférieurs à l'unité ont des logarithmes négatifs qui sont d'autant plus grands relativement que ces nombres sont plus petits, le logarithme de zéro étant l'infini négatif :

ANNÉES	NOMBRES ABSOLUS	LOGARITHMES	ANNÉES	NOMBRES ABSOLUS	LOGARITHMES
1885	0,40	$\overline{1}$.60205999	1901	2,33	0 36735592
1886	0,30	$\overline{1}$.47712125	1902	1,41	0.14921911
1887	0,48	$\overline{1}$.68124124	1903	1,23	0 08990511
1888	0,65	$\overline{1}$.81291336	1904	0,75	$\overline{1}$.87506126
1889	1,11	0.04532298	1905	0,83	$\overline{1}$.91907809
1890	28,4	0.45331834	1906	1 91	0 28103337
1891	1,82	0.26007139	1907	2,16	0.33445375
1892	0,60	$\overline{1}$.77815125	1908	1.37	0.13672057
1893	0 33	$\overline{1}$.51851394	1909	07,4	$\overline{1}$.86923172
1894	0 40	$\overline{1}$.60205999	1910	0.50	$\overline{1}$.69897000
1895	0 40	$\overline{1}$.60205999	1911	0.14	$\overline{1}$.14612804
1896	0 52	$\overline{1}$.71600334	1912	0,34	$\overline{1}$.53147892
1897	0,90	$\overline{1}$.95424251	1913	0.83	$\overline{1}$ 91907809
1898	1,05	0 02118930	1914	0.63	$\overline{1}$.79934055
1899	1,71	0 23299611	1915	0,75	$\overline{1}$.87506126
1900	4,26	0.62940960			

(1) Bowley, *Elements of statistics*, p. 189

Cherchant une division propre à établir l'échelle des ordonnées d'après une base logarithmique, nous observons que la différence entre les logarithmes de fr. 0.15 et 0.30 est : $\bar{1}.17609126 - \bar{1}.47712125$, soit .30103099 ou en abrégé, .301. Pour chaque intervalle logarithmique de .301 nous portons donc une longueur conventionnelle. Les ordonnées sont alors déterminées facilement et il ne reste qu'à procéder avec les logarithmes comme avec les nombres naturels, l'axe des abscisses étant, comme dans les diagrammes ordinaires, réservé aux divisions du temps.

Le principe essentiel de la courbe logarithmique est que les proportions des nombres entre eux restant égales, peu importent les nombres sur lesquels elles sont prises. Ainsi :

log. 0.15 = $\bar{1}.176$; log. 0.30 = $\bar{1}.477$ Différence = 301

mais

log 0.30 = $\bar{1}.477$; log. 0.60 = $\bar{1}.778$ Différence = .301

parce que ces logarithmes sont respectivement, dans les deux cas, dans le rapport de 1 à 2.

Du principe précédent, il résulte que les nombres naturels ne peuvent être disposés à intervalles égaux comme dans les diagrammes ordinaires, car la courbe logarithmique est basée sur les rapports et non sur les quantités absolues. Le diagramme basé sur ces principes reproduit, en courbe logarithmique, les données absolues exprimées graphiquement dans le diagramme simple. Bien que l'allure générale de la courbe ne soit pas beaucoup modifiée, il est à remarquer qu'aucune des données de la courbe logarithmique ne correspond exactement à la donnée correspondante du diagramme simple. En général, la forme est plus massive et les différences entre les sommets sont moins accentuées.

374. Mais l'utilité des courbes logarithmiques apparaît surtout quand il s'agit de traduire graphiquement une série ascendante dans laquelle les quantités absolues sont très différentes à la fin de ce qu'elles étaient au début. Dans une statistique relative, par exemple, aux dépôts existant dans

les caisses d'épargne, il est évident que le montant des dépôts est intéressant en lui-même : à cet égard, un historigramme ordinaire peut jouer un rôle utile. Mais cette figure ne peut nous apprendre qu'une chose, à savoir de combien de millions le montant total des dépôts a augmenté d'année en année. Mais pour l'économiste comme pour le sociologiste, il est une autre question beaucoup plus intéressante qui est celle-ci : quelles sont les époques auxquelles la classe des déposants a fait preuve de la plus grande capacité d'épargne? Cette question comporte une étude des nombres relatifs, et graphiquement elle ne trouve sa solution que dans la courbe logarithmique. Les données qui suivent constituent une application de la méthode aux données statistiques concernant les dépôts à la caisse générale d'épargne et de retraite en Belgique, de 1865 à 1900.

Montant des dépôts sur livrets par 100 habitants à la Caisse Générale d'Épargne et de Retraite de Belgique (matériel extrait de l'exposé de la situation du Royaume de 1876 à 1900, t. II, p. 696.)

Années.	Dépôts sur livrets par 100 habitants.	Logarithmes (1).	Années.	Dépôts sur livrets par 100 habitants.	Logarithmes (1).
1865	11	1.041	1883	2,302	3 362
1866	25	1.398	1884	2,869	3.410
1867	85	1.929	1885	3,047	3.484
1868	147	2.167	1886	3.475	3 541
1869	226	2 354	1887	3,831	3 583
1870	205	2.312	1888	4,146	3.618
1871	253	2 403	1889	4,483	3 652
1872	350	2 544	1890	5.198	3.716
1873	424	2.627	1891	5,273	3.722
1874	496	2.695	1892	5,452	3.736
1875	662	2 820	1893	6.036	3 781
1876	1,056	3.024	1894	6.526	3 815
1877	1,255	3.099	1895	6,861	3.836
1878	1.480	3 170	1896	7,227	3 859
1879	1,715	3.234	1897	7.848	3 895
1880	1,987	3 298	1898	8,298	3 919
1881	2,044	3.310	1899	8,846	3.949
1882	2,081	3.318	1900	9,669	3.985

(1) Il suffit de prendre les logarithmes avec 3 décimales, ce qui réalise une simplification notable.

Nous avons construit sur le même diagramme la courbe du diagramme orthogonal ordinaire et la courbe logarithmique.

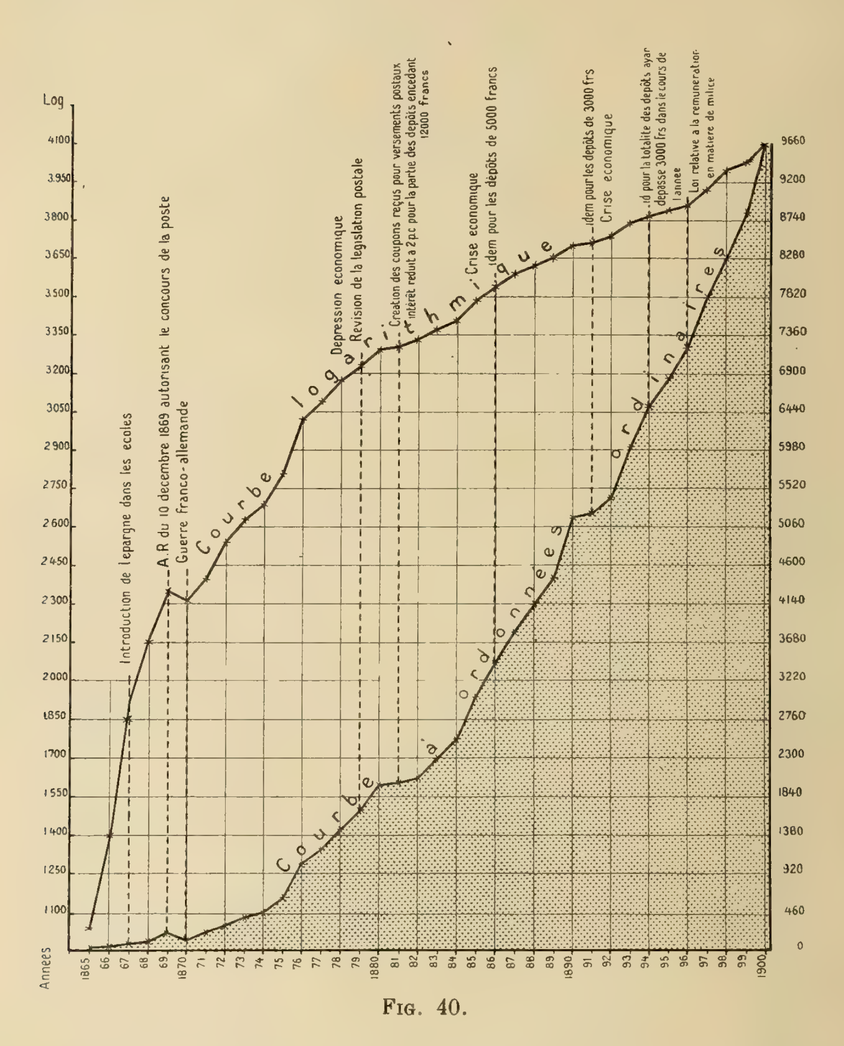


Fig. 40.

L'échelle des abcisses, portant les divisions du temps, est commune aux deux courbes. L'échelle des ordonnées pour le diagramme orthogonal simple est tracée de façon à ce que ses divisions soient en nombre égal à celles de l'échelle de la courbe logarithmique; chaque division est égale à

460 francs de sorte que toute augmentation de 1/10 sur l'échelle des ordonnées représente 46 fr. Pour l'échelle logarithmique, on a admis que toute division égale correspond à 150 fr. Les logarithmes étant ainsi localisés, on obtient une courbe très caractéristique, essentiellement différente de la courbe ordinaire. La réunion des deux courbes sur le même diagramme permet de lire immédiatement la valeur en francs équivalente à toute donnée logarithmique partant d'un point quelconque de cette courbe, il suffit de chercher le point situé sur la même ordonnée, appartenant à la courbe ordinaire et de lire en regard la valeur inscrite sur l'échelle de droite. Ainsi, en 1882, le point sur l'ordonnée logarithmique correspond à 1,840 francs plus 5 fois 46 francs = 2,070; la valeur réelle est 2,081, mais le diagramme ne permet pas une lecture d'une précision égale à celle d'une table numérique. Le procédé employé pour la construction de ce diagramme dispense du calcul d'une échelle de nombres absolus correspondant aux variations logarithmiques et présente encore l'avantage d'une comparaison immédiate entre les deux courbes. Le lecteur remarquera que sur la courbe logarithmique, l'inclinaison est beaucoup moins forte les dernières années qu'au début de la période observée. Les deux courbes sont intéressantes mais leur signification est différente : la courbe à ordonnées ordinaires traduit l'accumulation des sommes placées en dépôt en tenant compte du nombre d'habitants : la courbe à ordonnées logarithmiques marque l'effort réalisé par les épargnants proportionnellement aux efforts antérieurs; à mesure que les chiffres haussent, cette proportion a une tendance à diminuer parce qu'il est plus difficile d'atteindre un quantum d'augmentation sur de grands nombres que sur de petits. Au total, la courbe logarithmique marque un effort d'une remarquable continuité, mais très différent de celui qu'on pourrait déduire de la courbe ordinaire.

375. Les courbes logarithmiques comprennent une autre variété de courbes caractéristique auxquelles on a donné le nom de courbes à double échelle logarithmique. Elles sont aux premières ce que les diagrammes de distribution (polygones de fréquence) sont aux diagrammes de succession (historigrammes). Au lieu de porter sur l'axe des abscisses les divisions du temps, nous y portons une seconde échelle logarithmique basée sur les caractères de divisibilité du groupe : par exemple, les taux de salaires des ouvriers dont les nombres sont portés en ordonnées, ou encore, le taux du revenu des contribuables imposés d'après cette base. Dans les diagrammes à double échelle logarithmique, la courbe hyperbolique des diagrammes de distribution se change en une droite, comme il est facile de le démontrer.

D. — *Diagrammes polaires.*

376. Les diagrammes polaires sont basés sur une démonstration géométrique différente de celle qui s'applique aux diagrammes orthogonaux; il y a lieu de s'y arrêter un instant.

Sur un axe horizontal placé arbitrairement dans un plan, fixons un point fixe O considéré comme l'origine du système. Pour déterminer le point M situé dans le plan, il suffit de connaître la distance O M, quantité arithmétique que l'on désigne par la lettre p et l'angle positif $\omega < 2\pi$ compris entre p et l'axe horizontal. On appelle p le rayon vecteur du point M et ω l'angle polaire du même point; l'axe horizontal porte le nom d'axe polaire et le point fixe O est le pôle. La quantité p peut varier de zéro à l'infini positif et l'angle ω de 0 à 2π ; p et ω sont les coordonnées polaires de M.

On peut admettre des rayons vecteurs négatifs au-dessous de l'axe horizontal; on les porte sur la direction opposée à la direction M correspondant à l'angle ω. Dans ce cas, les coordonnées du point M_2 seront p et $2\pi - \omega$.

Toute relation entre p et ω représente en général une ligne, et réciproquement.

Dans les diagrammes polaires, les ordonnées sont portées autour d'un point central pour venir aboutir à l'axe des abcisses auquel on a donné la forme circulaire au lieu de le tracer dans le sens horizontal. Les ordonnées ont une longueur proportionnelle au fait qu'elles représentent; le fait est situé à une longueur M du centre, l'ordonnée est le rayon vecteur p. Les ordonnées sont en nombre variable; cependant, un trop grand nombre de données à représenter rendraient le diagramme polaire d'un usage difficile. On réserve généralement cette forme graphique aux phénomènes auxquels est liée l'idée de direction (direction des vents, etc.) ou aux faits qui se distribuent dans le temps selon un cycle invariable. L'axe des abcisses est mesuré par un rayon sur lequel sont portées les divisions quantitatives auxquelles on fait décrire un cercle. Les extrémités des ordonnées sont reliées entre elles par des droites. Les saillants montrent les quantités en excédent sur la valeur arbitraire, choisie pour abcisse, les rentrants marquent les quantités déficitaires.

L'exemple suivant peut servir d'illustration.

Nombre annuel de mariages par mois en Belgique, de 1896 à 1900

(Moyenne annuelle)

MOIS	NOMBRE DE MARIAGES	MOIS	NOMBRE DE MARIAGES
Janvier	4,174	Juillet	4,090
Février	4,966	Août	4,081
Mars	2,388	Septembre	5,146
Avril	6,045	Octobre	5,152
Mai	5,722	Novembre	4,842
Juin	4,365	Décembre	4,170

Total : 55,141. Moyenne : 4,595.

Pour construire ce diagramme, nous faisons la moyenne 4,595 égale à 1,000 et nous calculons les valeurs proportionnelles des données de chaque mois mais sur cette base. Représentant la moyenne par un rayon d'une longueur conventionnelle, nous traçons une circonférence représentant les valeurs de l'ordonnée moyenne; le cercle est divisé en 12 parties correspondant aux mois de l'année; la valeur relative de l'ordonnée est portée au centre de l'intervalle et tous les points sont ensuite reliés par des droites.

E. — *Cartogrammes.*

377. Ainsi que leur nom l'indique, les cartogrammes localisent sur une carte géographique les faits statistiques qu'ils expriment. Telle est la signification première du mot : cartogramme. Mais, par extension, le même nom a été donné à des représentations graphiques qui, sans l'aide des données géographiques, marquent l'intensité relative d'un fait au moyen de couleurs ou de parties ombrées, sans recourir aux mesures de surface ou de longueur qui sont le propre des diagrammes. Nous examinerons successivement ces diverses méthodes.

378. Parmi les cartogrammes proprement dits, nous distinguons : 1° les cartes avec diagrammes; 2° les cartes teintées; 3° les cartes à bandes; 4° les cartes avec courbes.

1° *Cartes avec diagrammes.* — On peut inscrire des diagrammes sur des cartes muettes de façon à localiser le fait qu'on veut représenter. L'élément essentiel est ici le diagramme, l'accessoire la carte géographique. Cela ne veut pas dire que l'élément géographique soit dépourvu de signification. On pourrait, il est vrai, se contenter d'écrire au-dessous de chaque diagramme le nom de la partie du territoire qu'il concerne, mais le fait de placer chaque graphique à l'endroit même de la carte où se passe le fait observé rend l'ensemble des constatations beaucoup plus clair et plus suggestif. Qu'il s'agisse de la culture du froment ou de

la vigne, de la proportion des naissances ou des décès par 1,000 habitants, du nombre de crimes, ou du mouvement des ports, l'utilisation d'une carte est un procédé d'une grande utilité : d'abord, parce qu'il groupe les éléments à étudier et attire forcément l'attention sur leur concentration ou leur dispersion; ensuite parce qu'il suggère l'étude des relations du phénomène avec le milieu. On réussit par ce moyen à donner une idée claire de phénomènes très complexes. M. Cheysson, par exemple, a représenté l'accroissement ou la diminution de la population dans chaque département français à chacun des quatorze recensements qui avaient été exécutés à l'époque où son travail a paru : il l'a fait au moyen de cercles divisés en quatorze secteurs et placés chacun sur le territoire du département. Quand on a à comparer les résultats d'un dénombrement à deux époques, on peut recourir à l'emploi de rectangles d'une grandeur proportionnelle au fait, divisés en deux parties dont l'une, colorée, marque l'intensité du phénomène au moment de la dernière observation et dont l'autre, ombrée, signifie l'intensité à l'époque antérieure.

379. Les graphiques combinés avec les cartogrammes sont un moyen précieux de réaliser la représentation de phénomènes complexes. En 1896, le recensement général de l'industrie et des métiers qui fut fait à cette date en Belgique, nous a donné l'occasion de présenter sous une forme graphique, simple et rapide, le phénomène complexe de l'attraction exercée par les centres industriels sur les localités environnantes; il s'agissait de montrer, par un cartogramme intuitif : 1° les communes dont se composait le centre industriel étudié; 2° les communes où habitaient les ouvriers allant travailler dans ce centre industriel; 3° le nombre de ces ouvriers (approximativement) ; 4° le groupe industriel duquel dépendait leur profession. Après avoir défini le sens donné à l'expression « centre industriel » nous avons tracé sur une carte géographique le contour des

communes composant ce centre ; tout autour sont disposées les communes dont une certaine population industrielle va travailler dans ce centre. Pour le nombre des ouvriers et le groupe auquel ils appartiennent, on a commencé par attribuer un numéro d'ordre à chaque groupe de profession (au total 8) ; la présence d'ouvriers appartenant à un certain groupe industriel est signalée pour chaque commune, par l'inscription sur le territoire de cette commune d'un petit carré ([ ]) à l'intérieur duquel se trouve un chiffre correspondant au numéro d'ordre du groupe. Les chiffres droits [2] représentent une dizaine, les chiffres inclinés [*2*] une fraction de dizaine d'ouvriers appartenant à l'industrie à laquelle ce chiffre est attribué. Ainsi [1] [1] [1] [*1*] signifie : entre 30 et 40 ouvriers appartenant à l'industrie des mines, habitant la commune où ces chiffres sont inscrits et allant travailler dans l'agglomération urbaine ou industrielle à laquelle la carte se rapporte (1).

380. *Cartes teintées.* — Dans les travaux statistiques, beaucoup de données sont publiées d'après une base géographique ou politique : la région agricole ou industrielle, la province, département ou comté, le district, le canton, etc. Il était dès lors tout indiqué d'employer des cartes pour localiser l'expression du phénomène, mais au lieu de recourir, pour ce faire, à des diagrammes, on a eu l'idée d'employer les couleurs pour signifier l'importance relative des résultats. Bien que le procédé soit très simple et presque dépourvu de difficultés, il n'en est pas moins vrai que des bévues se commettent encore : ainsi le manque de relation entre l'unité géographique choisie et le fait spécial à représenter. Dans un recensement agricole, on a, par exemple, adopté pour unité géographique de représentation graphique... le canton judiciaire ! C'est la même chose que si,

---

(1) Royaume de Belgique. *Atlas statistique du recensement général des industries et des métiers* (31 octobre 1896), Bruxelles, 1903, p. 5. (Cfr. Planches XVI à XXIII).

pour la statistique judiciaire, on prenait pour base géographique le ressort des agronomes de l'Etat. A défaut de division naturelle adéquate, il faut se borner à employer les divisions politiques usuelles. De ces dernières, la meilleure est la commune parce qu'elle est la plus homogène. Dans les grands pays, la présentation graphique ou statistique, par commune, est parfois irréalisable, mais, dans les petits pays, elle est fort à recommander.

L'établissement d'une carte teintée repose sur ce principe : étant donnée une collectivité, un phénomène complexe dont les manifestations sont plus ou moins intenses, le statisticien opère des groupements, d'après lesquels il arrête une échelle de teintes : à la teinte la plus claire correspond la densité la plus faible et ainsi de suite. L'échelle des faits peut être établie en tenant compte des nombres absolus ou des nombres proportionnels ou relatifs. On peut aussi rapporter les chiffres absolus à une unité convenue. Ainsi, l'échelle des faits peut être déterminée dans un recensement de la population, d'après le nombre d'habitants, par kilomètre carré; dans le recensement allemand du 14 juin 1895, nous voyons, par exemple, que la densité de la population est exprimée au moyen de sept catégories (de 15.68 hab. à 2,036 hab. par kilomètre carré) auxquelles correspondent autant de teintes distinctes obtenues dans une seule couleur au moyen de « grisés » (1). Mais, en général, l'emploi des nombres relatifs doit être préféré : ce procédé est plus exactement représentatif que l'autre, en ce qu'il élimine les éléments variables qui peuvent modifier la répartition géographique du fait envisagé, et parce qu'il facilite ainsi les comparaisons.

La distribution des entreprises de l'industrie alimentaire, en nombre absolu, donne une grande importance à tous les centres urbains; la base est meilleure si l'on consi-

---

(1) *Die berufliche und soziale gliederung des Deutschen Volkes nach der berufzählung von* 14 *Juni* 1895, *Berlin*, 1899, p. 11.

dère le nombre de ces entreprises pour 10,000 habitants, l'influence des grands centres se trouvant ainsi annihilée (1).

381. Des discussions se sont engagées à propos du nombre de couleurs à employer : certains spécialistes ne veulent entendre parler que de teintes monochromes, d'autres sont résolument polychromistes. Ces derniers font observer qu'il est difficile d'arriver à exprimer suffisamment de rapports de variabilité à l'aide d'une couleur unique, dont les teintes, au delà d'un nombre d'ordinaire limité, ne peuvent que difficilement se différencier. On peut répondre que l'échelle de variabilité peut être assez restreinte sans nuire, en général, à l'exactitude de la présentation statistique et que, somme toute, par l'emploi de « grisés », on arrive à multiplier le nombre de teintes qu'on peut tirer d'une couleur unique : pour notre part, nous en avons obtenu huit ou neuf, très aisément perceptibles, ce qui est largement suffisant dans le plus grand nombre de cas (2).

M. Levasseur, souvent, emploie deux couleurs : le rouge et le bleu ; le rouge pour toutes les divisions territoriales où le fait s'élève au-dessus de la moyenne de l'ensemble du pays, le bleu pour celles où le fait est inférieur à la moyenne. Ce système a donné à l'auteur de bons résultats. Au surplus, la controverse est dénuée de portée scientifique si l'on admet, de part et d'autre, que la qualité maîtresse d'une représentation graphique consiste dans sa clarté.

Les couleurs peuvent être remplacées par des hachures imprimées en noir. Ce système est beaucoup moins coûteux que l'impression en couleurs, mais le résultat est moins attrayant et ne présente pas autant de clarté.

La publication des cartogrammes est devenue plus abor-

---

(1) March (L.), « Les représentations graphiques et la statistique comparative ». (*Journal de la Société de Statistique de Paris*, 1905.)

(2) Voyez *Atlas statistique du recensement général des industries et des métiers*, Bruxelles, 1903.

dable, depuis qu'on a eu l'idée de schématiser les cartes géographiques en divisant la surface du pays en un grand nombre de parties se coupant à angles droits, correspondant aux divisions géographiques, s'emboîtant les unes dans les autres comme un jeu de puzzle; on a appliqué le même système aux divisions politiques pour figurer le nombre d'élus de chaque parti dans les divers arrondissements électoraux. La carte en couleurs de la composition du Reichstag allemand après les élections de 1903, reproduite à la suite du rapport de M. Mayet (1) est extrêmement intéressante; en lithographie, on évite de la sorte un laborieux travail de dessin et le repérage est plus facile; en impression ordinaire, l'imprimeur prépare à l'avance, pour chaque division administrative, la gamme complète des clairs et des ombres, de façon qu'il dispose à l'avance d'un matériel adéquat à tous les degrés de variabilité du phénomène. Les maisons d'impression se chargent même d'opérer la division schématique du territoire d'après les indications du bureau de statistique.

382. *Cartes à bandes.* — Une application ingénieuse de la méthode graphique à la statistique des transports a été l'invention des cartes à bandes. Des bandes coloriées, tracées sur une carte géographique, suivent la direction des voies de communication; elles sont d'une largeur proportionnelle à l'importance du fait qu'elle représentent. Une même bande peut se composer de plusieurs bandes parallèles, comme un total peut se composer de plusieurs éléments. Ces diagrammes ont l'avantage de réunir les renseignements portant sur les quantités et rendent de grands services dans la statistique des transports : chemins de fer, canaux, routes, ports en rivière et ports de mer. Leur calcul ne présente aucune espèce de difficulté : on adopte au début

(1) MAYET, « Die schematisch-statistichen Karten des Kaiserlichen statistichen Amtes zu Berlin ». (*Bulletin de l'Institut International de Statistique*, t. XIV, 3e livre, p. 214.)

la convention qu'un certain nombre de tonnes de marchandises ou un nombre de voyageurs sera représenté par une bande d'une largeur déterminée. On peut distinguer les transports d'après le genre de moyen de locomotion en se servant de couleurs ou de hachures. Le lecteur trouvera de nombreuses applications de ces diagrammes dans les « albums de statistique graphique » publiés autrefois par M. Cheysson.

383. *Cartes avec niveau.* — Dans les cartes géographiques, les topographes emploient, pour figurer le relief et les dépressions du terrain, les courbes de niveau : ce sont des plans horizontaux, équidistants et parallèles, qui passant par un grand nombre de points dont l'altitude est connue, figurent l'aspect du sol avec ses collines, ses vallées et ses plaines. Lorsque les courbes de niveau sont nombreuses et serrées les unes contre les autres, c'est une colline ; lorsqu'elles sont largement espacées, c'est une plaine que nous avons sous les yeux. La même convention peut être adoptée à l'égard des faits statistiques : au lieu de collines, admettons qu'il s'agisse de populations fortement agglomérées (plusieurs centaines par kilomètre carré) ; nous supposerons que les faits sont superposés les uns aux autres et par tous les points d'une égale importance, entre les limites de chaque classe, nous ferons passer une courbe de niveau. On obtient de la sorte une carte statistique avec niveau, à la condition que pour chaque point du territoire (la commune), on dispose des renseignements de fait nécessaires. Le procédé est laborieux, d'une exécution lente et d'une lecture assez malaisée ; nous lui préférons, sans aucun doute, les simples cartes teintées. Une remarque analogue s'applique aux cartes en relief ; M. Levasseur en a fait une critique qui nous semble justifiée (1).

(1) LEVASSEUR, « La Statistique graphique ». (*Jubilee volume of the Stat. Soc.*, p. 245.)

384. *Autres figures coloriées* (non accompagnées de cartes). — Bien que les cartogrammes soient, par essence, des représentations statistiques au moyen de cartes, on range dans la même catégorie, les figures coloriées au moyen desquelles se marquent les rapports de densité. On en a fait quelques applications heureuses que nous signalons brièvement.

Un maître regretté de la statistique française, M. de Foville, s'est efforcé de marquer les variations d'un certain nombre d'éléments intéressant la prospérité publique, au moyen de teintes et de couleurs (1). Le système graphique de M. de Foville consiste, comme l'auteur l'explique lui-même, en une sorte de table de Pythagore, plus haute que large, comprenant dans le sens de la hauteur autant de bandes horizontales que d'indices consultés et, dans le sens de la largeur, un nombre de bandes verticales aussi fréquentes que les années sur lesquelles s'étend la recherche. Selon que les résultats sont qualifiés de bons, assez bons, médiocres ou mauvais, les carrés formés par l'intersection des bandes verticales et horizontales sont coloriés en rouge, rose, gris ou noir. Il en résulte une impression d'ensemble de nature intuitive. Ce système n'échappe pas à quelques critiques ; nous en parlerons au tome second de cet ouvrage, où nous examinerons avec détail les procédés de la « Sémiologie économique ».

M. Benini a donné dans ses « Principii di statistica » un exemple intéressant de cartogramme non géographique, ainsi qu'il dénomme cette figure (2). Le problème posé est le suivant : pour onze groupes d'âges divisant la population italienne en 1886, exprimer graphiquement la mortalité relative par mois et par groupe d'âge. L'auteur a commencé par établir sept gradations de mortalité, le nombre

---

(1) de Foville, « Essai de Météorologie économique et sociale ». (*Journal de la Société de Statistique de Paris*, 1888, p. 243.)

(2) Benini, *loc. cit.*, pp. 154-155.

sept ne pouvant guère être dépassé si l'on veut obtenir des teintes faciles à distinguer les unes des autres; le chiffre relatif le plus bas renseigné dans la table originaire est 71; le plus élevé est 151; la différence 80 divisée par 7 donne pour quotient 11.43; chaque classe se forme par l'addition de 11.43 au chiffre le plus bas, puis au chiffre suivant immédiatement le chiffre le plus élevé du groupe précédent et ainsi de suite.

Onze bandes verticales représentent les onze catégories d'âges; chacune est divisée en 12 parties égales correspondant à un mois de l'année. Le résultat est très frappant en ce sens qu'il marque immédiatement l'existence d'un maximum de mortalité, pour les classes d'âges les moins élevées, pendant les mois chauds (juillet, août, septembre); cette mortalité diminue progressivement à mesure qu'on s'élève dans l'échelle des âges, sauf que le mois d'août reste longtemps celui où la mortalité est la plus forte; pour les vieillards, le maximum de la mortalité s'observe au contraire en hiver. (Décembre à mars.)

F. — *Stéréogrammes.*

385. Toutes les figures que nous avons passées en revue, étant situées dans un plan, deux éléments seulement étaient à déterminer; les diagrammes solides ou stéréogrammes sont des figures situées dans l'espace et, par conséquent, un troisième élément s'ajoute aux deux autres : cette circonstance, tout en compliquant le graphique, le rend apte à figurer les phénomènes statistiques qu'il serait impossible de représenter dans un plan. Les diagrammes ordinaires ne permettent que la représentation de deux faits, comme le nombre de décès et le temps, le nombre des ouvriers gagnant un salaire et le taux de ce salaire, la taille des conscrits et leur nombre par grandeur. Le plus souvent, le but de la statistique est atteint quand elle dispose de deux éléments, parfois elle en réclame davantage; c'est alors

qu'apparaît l'utilité du stéréogramme qui vient combler une réelle lacune ; ainsi, par exemple, on peut mettre en relation le nombre de naissances avec l'âge du père et de la mère au moment du mariage, etc. Le statisticien italien Perozzo, qui avait été chargé par la direction générale de la statistique du royaume d'Italie de dresser des modèles de stéréogrammes, en a fait plusieurs qui sont restés classiques et dont on trouvera la reproduction et la description dans un grand nombre de traités, ce qui nous dispense d'en reproduire le dessin avec les commentaires. Nous notons seulement les éléments qui se peuvent distinguer sur le stéréogramme construit par le savant géomètre, relatif à la population de la Suède de 1750 à 1875 ; leur nombre est la démonstration la plus évidente de la supériorité du stéréogramme sur le diagramme comme quantité de faits emmagasinés. Ce sont : 1° le mouvement général des naissances ; 2° les variations du chiffre de la population à chaque recensement ; 3° le nombre des recensés à chaque âge ; 4° le nombre de survivants à chaque génération ; 5° les individus en force numérique égale. Les crises économiques et sociales peuvent aussi être suivies sur le stéréogramme, comme on peut y voir, vingt-cinq ou trente ans plus tard, les conséquences de la diminution des naissances.

S'il est exact de dire qu'aucune autre forme de diagramme n'est capable de réunir un aussi grand nombre de données, il ne faut point manquer d'ajouter que la représentation simultanée d'éléments aussi nombreux conduit à la complication et à la confusion. Un diagramme perd la plus grande partie de son utilité si la clarté lui fait défaut.

Malgré ce que peut avoir d'un peu effrayant, à première vue, l'emploi des trois dimensions, la construction du stéréogramme ne donne pas lieu à de très grosses difficultés. Nous essayerons d'en déterminer les principes d'après un exemple simple et en éliminant de notre exposé tout ce qui pourrait rebuter le lecteur.

Nous avons vu précédemment ce qu'on entend par « table de corrélation ». C'est un tableau à double entrée dans lequel les fréquences sont indiquées par paires d'éléments. Un exemple de table de corrélation est celle qui est formée des données relatives à l'âge du mari et à l'âge de la femme au moment de la célébration du mariage, à une certaine année. Nous avons précisément donné un exemple semblable emprunté à la statistique de la Belgique. (Cfr. n° 318.) Un autre exemple de table de corrélation est celui du nombre d'enfants d'une femme, et le nombre d'enfants qu'a eus la fille de cette femme. Dans le sens horizontal sont inscrites les données relatives au nombre d'enfants des mères et dans le sens vertical, le nombre d'enfants des filles (1).

Nous pouvons, à l'aide d'une table de corrélation, décrire aisément la construction d'un stéréogramme. Nous supposons qu'on trace sur une surface plane une série de traits correspondant à l'emplacement de chaque colonne : traits verticaux pour les âges du mari portés en abcisses, traits horizontaux pour les âges de la femme, portés en ordonnées. On obtiendra ainsi une sorte de surface réticulaire; à l'intersection des lignes, on pourra lire certains chiffres indiquant le nombre de mariages accomplis entre des hommes de tel âge avec des femmes de tel âge. Qu'on élève ensuite, à ce point de la surface, une ligne verticale dont la hauteur correspondra au nombre des unités indiqué au point de rencontre et qu'on réunisse tous les sommets des verticales : on obtiendra de la sorte un stéréogramme qu'on appelle une surface de fréquence, qui correspond aux courbes ordinaires dont il a été parlé au chapitre de la distribution. Si la distribution des fréquences est idéale, c'est-à-dire que pour tous les âges elle commence par des nombres peu élevés, pour atteindre

(1) Mémoire de K. Pearson, Alice Lee et L. Bramley Moore. Phil. Trans., vol. CXCII, 1899, tableau IV.

son maximum vers la moyenne et diminuer ensuite graduellement, la surface de fréquence aura une forme régulière, de même que la courbe idéale de fréquence des diagrammes ordinaires. Dans une surface de fréquence absolument régulière, une des sections quelconques du solide reproduirait la figure de la courbe binomiale régulière. Nous avons vu que les courbes peuvent se subdiviser, comme Pearson l'a montré, en un certain nombre de types auxquels toutes les courbes quelconques peuvent être ramenées. Mais il n'en est pas de même pour les surfaces de fréquence parce que leurs variétés sont trop nombreuses ; la symétrie parfaite dans les surfaces de fréquence est encore plus rare que dans les courbes de fréquence.

Le lecteur remarquera que cette construction n'est qu'une extension de la méthode cartésienne des coordonnées. Au lieu de mesurer au moyen de deux axes orthogonaux, x et y, les valeurs de deux variables, on emploie ici trois axes, x, y et z orthogonaux entre eux, à l'effet de mesurer trois variables, deux axes x et y restent dans un plan horizontal, tandis que le troisième est dressé dans la position verticale ; les constructions de l'espèce dépendent de la géométrie analytique dans l'espace.

386. Les stéréogrammes présentent un intérêt scientifique et théorique qui n'a pas son pendant, sous le rapport pratique. Ces figures sont à la fois trop coûteuses et trop encombrantes, pour l'enseignement, si elles sont faites en relief. Si on les remplace par une figure géométrique, elles ne sont utiles qu'à ceux qui possèdent des connaissances plus ou moins étendues en géométrie, et précisément, parmi les personnes s'appliquant à l'étude des questions économiques et sociales, ces connaissances ne sont pas des plus répandues. C'est ici le lieu de rappeler ces paroles de M. Levasseur : « Il ne faut pas oublier que les graphiques statistiques sont surtout un moyen de vulgariser les nombres, et que, dès qu'ils demandent à l'esprit plus d'efforts

qu'il n'en faut pour étudier et comparer ces nombres, ils n'ont plus de raison d'être (1) ».

III. — La statistique graphique comparative.

387. La fonction des graphiques n'est pas seulement démonstrative; elle est, tout autant, comparative. L'idée de représenter une foule de chiffres par une simple courbe épousant les variations de la série et en traçant le schéma d'une façon rapide et sûre, fait naître nécessairement cette autre idée d'employer le même procédé à la comparaison. S'il n'y avait que peu de nombres à rapprocher les uns des autres, on n'éprouverait pas le besoin de recourir aux graphiques; mais on se trouve d'ordinaire devant des dizaines, parfois des centaines de chiffres à comparer... Pour suppléer à l'insuffisance de notre mémoire, nous avons heureusement les graphiques, dont nous pouvons, sous certaines conditions, rapprocher les renseignements. On peut ainsi définir le rôle des représentations graphiques sous le rapport comparatif : 1° elles mettent en lumière les relations réciproques des phénomènes; 2° elles peuvent aider à découvrir ce qu'il y a de permanent dans ces relations; 3° elles servent à l'étude des phases d'un même phénomène et en décèlent les irrégularités et les anomalies; 4° elles peuvent aider à rectifier les données fautives et à compléter les séries où se remarque quelque lacune (2).

388. Suffit-il, pour obtenir une vue rapide et précise des points à comparer, de rapprocher deux diagrammes quelconques? On sent très bien que non. Si la relation entre l'unité de mesure des abcisses et l'unité de mesure des ordonnées n'est pas la même dans les diagrammes qu'on dé-

---

(1) Levasseur, « La statistique graphique ». (*Jubilee volume of the Statist. Soc.*, London, 1885, p. 247).

(2) March (L.), « Les représentations graphiques et la statistique comparative ». (*Bulletin de la Société de Statistique de Paris*, 1904, p. 408).

sire comparer, il n'y a pas lieu de les rapprocher. Les graphiques peuvent servir, en des mains malhonnêtes, à créer des fantasmagories auxquelles se laissent prendre trop souvent les non-initiés et les naïfs. Avant donc de comparer un tracé graphique à un autre, il faut s'assurer qu'ils sont bien l'un et l'autre à la même échelle et s'il n'en est pas ainsi il faut employer une formule de réduction très simple, recommandée par l'Institut international de statistique (1) : si l'échelle des diagrammes n'est pas la même, il suffit, pour les unifier, de substituer aux nombres à représenter le rapport de toutes les ordonnées à leur valeur moyenne calculée sur une période fixe, par exemple 1901-1910 et dont la grandeur de base, égale à 100, représentera sur l'axe des ordonnées une distance égale à celle qui, sur l'axe des abscisses, mesure trente années.

A l'aide de la réduction arithmétique qui vient d'être indiquée, on peut très facilement et avec certitude comparer entre elles des courbes chronologiques.

389. On vient de voir que pour comparer entre elles des courbes construites à des échelles différentes, on est forcé de les ramener à des chiffres proportionnels. L'avantage de la méthode, qui est la possibilité de comparer entre eux des éléments qui, auparavant, n'étaient pas comparables, est évident; il ne va pas cependant sans un léger inconvénient : le lecteur se trouve dépourvu des données en nombres absolus qui donnent leur valeur aux chiffres proportionnels. Pour remédier à cette lacune, il suffit de placer en regard de l'échelle des chiffres proportionnels une seconde échelle consacrée aux chiffres absolus : prenant un point quelconque de l'échelle proportionnelle, on calcule sa valeur en nombres absolus et on divise la hauteur totale de

(1) Cfr. *Bull. Instit. Intern. Stat.*, t. XIX, première livraison, p. 118 et suivantes. Voyez également MARCH (L.), *loc. cit.*, pp. 415-416.

l'échelle en divisions égales correspondant au chiffre de base.

L'ensemble des règles indiquées ci-dessus est de stricte interprétation; à défaut de s'y conformer il n'est pas permis de comparer entre eux des diagrammes chronologiques ou de succession.

390. La comparaison des courbes de distribution doit être envisagée séparément, à raison de particularités qui lui sont propres.

Pour obtenir une comparaison rationnelle entre deux courbes de distribution, il faut non pas comparer entre eux les nombres absolus, mais il convient de figurer au moyen du tracé graphique, les nombres proportionnels. On se rend très bien compte du jugement instinctif, mais faux, qui serait porté par le lecteur d'un graphique basé sur les nombres absolus : supposons comme M. March l'a fait, deux catégories d'ouvriers A et B ayant exactement la même distribution de salaires, mais dont l'une (B) est deux fois plus nombreuse que l'autre (A); traçons les deux courbes, en élevant les ordonnées à une hauteur proportionnelle au nombre d'ouvriers par classes de salaires, il en résulte que la courbe *b* aura des ordonnées d'une longueur double à celle de toute ordonnée quelconque de *a* et que la proportion entre sa hauteur et sa base sera toute différente de celle qui existe pour *a*. On devrait donc conclure que la distribution des salaires en B n'est pas la même que la répartition qui se remarque en A, ce qui est faux par hypothèse. La comparaison au moyen de nombres proportionnels est donc une nécessité en ce qui concerne les courbes de distribution.

La distribution comparée des salaires parmi les ouvriers tisserands (à la mécanique) de jute, de lin, de coton et de laine s'établit d'après les règles précédentes en se basant sur les données numériques ci-après :

Distribution des salaires parmi les tisserands mâles de plus de 16 ans, en octobre 1901

(Matériel extrait de : *Salaires et durée du travail dans les industries textiles au mois d'octobre 1901, Bruxelles, 1905, p. 154 et 155, analyse*.)

Textiles	Nombre des ouvriers de plus de 16 ans ayant gagné par jour																		
	moins de 0.50	0.50 à 0.99	1.00 à 1.49	1.50 à 1.99	2.00 à 2.49	2.50 à 2.99	3.00 à 3.49	3.50 à 3.99	4.00 à 4.49	4.50 à 4.99	5.00 à 5.49	5.50 à 5.99	6.00 à 6.49	6.50 à 6.99	7.00 à 7.49	7.50 à 7.99	8.00 à 8.49	8.50 à 8.99	Tota
Jute . .	—	9	103	111	168	122	14	3	2	1	—	—	—	—	-	—	—	—	553
Lin . . .	—	4	203	539	872	532	353	174	50	13	2	—	—	—		-	-	—	2742
Coton . .	—	4	103	352	499	517	420	280	171	70	21	5	1	—		—	—	—	2443
Laine . .	—	3	82	257	449	536	508	389	276	199	81	45	32	14	—	-	—	—	2871

30
25
20
15
10
5
0

moins de 50 c
de fr 0.50 à fr 0.99
de fr 1.00 à fr 1.49
de fr 1.50 à fr 1.99
de fr 2.00 à fr 2.49
de fr 2.50 à fr 2.99
de fr 3.00 à fr 3.49
de fr 3.50 à fr 3.99
de fr 4.00 à fr 4.49
de fr 4.50 à fr 4.99
de fr 5.00 à fr 5.49
de fr 5.50 à fr 5.99
de fr 6.00 à fr 6.49
de fr 6.50 à fr 6.99
de fr 7.00 à fr 7.49
de fr 7.50 à fr 7.99
de fr 8.00 à fr 8.49
fr 8.50 et plus

Fig. 41.

Les diagrammes comparatifs dressés d'après cette méthode sont très clairs et tiennent compte de toutes les exigences théoriques de la comparaison : on voit parfaitement sur le diagramme, combien la distribution des salaires varie d'après le textile mis en œuvre, encore que la spécialité professionnelle (tisserands), le sexe, l'âge et l'outillage (tissage mécanique) soient les mêmes dans tous les cas ; on ne pourrait souhaiter de démonstration plus évidente et répondant plus exactement aux exigences de la comparaison.

On pourrait cependant regretter que la substitution des nombres proportionnels aux chiffres absolus ait pour effet de mettre toutes les données sur le même plan, malgré que leur importance respective soit fort différente : la courbe de 533 tisserands de jute a autant d'importance que celle relative aux 2,871 tisserands de laine. Il est facile de remédier à cet inconvénient, quand il y a deux courbes à comparer, en reproduisant l'échelle des nombres absolus du côté opposé à celui occupé par l'échelle des nombres relatifs, mais quand il y a un certain nombre de courbes portées sur la même feuille, il faut nécessairement isoler les courbes dont on veut retracer les variations absolues et relatives.

Un procédé élégant est ainsi indiqué par M. L. March (1) : le rapport entre les nombres absolus étant connu, par exemple, 2.25, il suffit, pour donner aux courbes proportionnelles une hauteur répondant aux nombres absolus, de prendre pour unité des abcisses une longueur égale à l'unité d'abcisse du diagramme comparatif multipliée par la racine carrée du rapport 2.25, et de procéder de même pour l'unité des ordonnées. « Avec ces unités, l'impression des changements absolus et relatifs est exacte, et, au moyen d'une double échelle, on peut aussi bien mesurer les uns que les autres. »

---

(1) MARCH (L.), *loc. cit.*, p. 419.

391. La statistique comparative trouve son application usuelle dans les diagrammes, mais ce serait une erreur de penser qu'elle doive rester étrangère à la matière des cartogrammes. M. Cheysson a exposé, en 1887, les moyens de rendre comparables entre elles les diverses cartes d'une série de cartogrammes, et son système peut être indiqué et suivi aujourd'hui encore (1).

M. Cheysson a fait remarquer qu'il existe un grave inconvénient à traiter chaque cartogramme au mieux de ses nécessités graphiques et de ses particularités individuelles, sans se préoccuper de l'ensemble, lorsque la série des cartogrammes se rapporte à des faits homogènes et de même famille. Cet inconvénient consiste en ce que des cartogrammes qui traduisent aux yeux la même opération sous ses divers aspects ne peuvent, sans égarer le lecteur, donner la même importance graphique à des faits de très inégale densité. La classe professionnelle des médecins, partagée en huit teintes, réparties entre les provinces ou les arrondissements d'un Etat, apparaîtra aussi dense que la catégorie des agriculteurs, également divisée en huit teintes; cependant, la teinte la plus foncée correspondra en fait à un bien plus grand nombre d'agriculteurs que de médecins... Pour éviter ce trompe-l'œil, comment faut-il procéder? La substitution de nombres proportionnels aux nombres absolus ne peut ici obvier en rien à l'inconvénient signalé qui consiste dans l'attribution d'une même teinte occupant le même rang dans l'échelle des teintes dégradées, à deux faits de densité différente, que cette densité soit exprimée par des chiffres absolus ou par des chiffres proportionnels. M. Cheysson a proposé de remplacer les coefficients absolus par leurs écarts par rapport à la moyenne générale.

---

(1) CHEYSSON (E.), « Les cartogrammes à teintes graduées », (*Journal de la Société de Statistique de Paris,* 1887, p. 128).

La formule de l'écart proportionnel sera $e = \frac{d - m}{m}$ dans laquelle d est l'écart de chaque coefficient et m la moyenne générale.

Comme l'a fait remarquer l'auteur à qui est due la méthode, la substitution des écarts aux chiffres absolus conserve l'importance respective des faits, leur hiérarchie, et permet de retrouver sans peine le fait lui-même sans l'écart. Une même teinte appliquée à des faits homogènes ne signifiera donc pas qu'ils ont une densité comprise dans les mêmes limites, mais simplement que l'écart qu'ils présentent respectivement à la moyenne est le même.

Il est évident, comme le dit encore M. Cheysson, que l'étude d'un atlas dont toutes les cartes seraient basées sur ce principe, est plus instructive que celle de documents graphiques ayant chacun leur signification propre et ne souffrant de comparaison avec aucun autre.

IV. – La statistique graphique comme instrument d'investigation.

392. La statistique graphique peut aussi servir d'instrument d'investigation et se substituer au calcul. Nous passerons en revue quelques-unes des principales applications qui ont été faites de cette propriété des graphiques.

A. — *Méthode graphique des pourcentiles de Sir F. Galton.*

Dans son ouvrage « Natural inheritance », Sir F. Galton a exposé une méthode qu'il appelle « méthode des percentiles », pour arriver rapidement à la détermination graphique des quartiles et de la médiane. Bien que, sous certain aspect, ce procédé se rattache directement à la mesure de dispersion interquartile, la méthode de Galton ne peut être confondue avec la matière exposée sous le n° 287, à raison de la portée plus générale que lui a donnée son illustre

auteur : les services que peut rendre la courbe galtonienne vont bien au delà, en effet, de la détermination de la médiane.

Pour construire la courbe de Galton, on procède de la sorte :

On commence par dresser le tableau numérique des données à utiliser, par exemple, les salaires des 10,455 ouvriers (mâles) de plus de 16 ans, occupés en octobre 1903, en Belgique, dans l'industrie de la construction de machines motrices, machines-outils et appareils industriels.

Pour que la courbe soit tout à fait exacte, il faut que la série statistique soit complète, c'est-à-dire qu'elle ne comprenne pas de classe initiale ou terminale en renfermant plusieurs qui restent indéterminées.

Dans la première colonne, on porte le taux des salaires, dans la deuxième le nombre d'ouvriers gagnant le salaire indiqué en regard, dans la troisième la proportion de chaque groupe au total, dans la quatrième on additionne l'un à l'autre les pourcentages inscrits à la troisième colonne en marquant chaque fois le résultat de l'addition.

Le tableau ci-après contient les données statistiques et le résultat des calculs en ce qui concerne les 10,455 ouvriers de l'industrie de la construction de machines dont les salaires ont été relevés en 1903 :

Salaires de 10,455 ouvriers mâles de plus de 16 ans dans l'industrie de la construction de machines motrices, machines-outils et appareils industriels en Belgique (*matériel extrait de « Salaires et durée du travail dans les Industries des métaux au mois d'octobre 1903 »*. Publication de l'office du travail de Belgique, Bruxelles. 1907).

Salaires.	Nombre d'ouvriers.	Pourcentages.	Pourcentages accumulés.
de 0.25 fr. à 0.49 fr.	1	0,009	0,009
» 0.50 » 0.74	14	0,133	0,142
» 0.75 » 0.99	12	0,114	0,256
» 1.00 » 1.24	91	0,874	1,130
» 1.25 » 1.49	106	1,013	2,143
» 1.50 » 1.74	182	1,740	3,883
» 1.75 » 1.99	172	1,645	5,528
» 2.00 » 2.24	289	2,764	8,292
» 2.25 » 2.49	285	2,725	11,017
» 2 50 » 2.74	471	4,505	15,522
» 2.75 » 2 99	555	5,308	20,830
» 3.00 » 3.24	957	9,153	29,983
» 3.25 » 3.49	700	6,695	36,678
» 3.50 » 3,74	1029	9,842	46,520
» 3.75 » 3.99	926	8,857	55,377
» 4.00 » 4.24	1209	11,563	66,940
» 4.25 » 4.49	729	6,972	73,912
» 4.50 » 4.74	821	7,852	81,764
» 4.75 » 4.99	470	4,495	86,259
» 5 00 » 5 24	535	5,117	91,376
» 5.25 » 5 49	183	1,750	93,126
» 5 50 » 5.74	270	2,582	95,708
» 5.75 » 5 99	103	0,985	96,693
» 6.00 » 6 24	121	1,157	97,850
» 6.25 » 6 49	42	0,401	98,251
» 6.50 » 6.74	60	0,573	98,824
» 6.75 » 6.99	20	0,191	99,015
» 7.00 » 7.24	31	0,296	99,311
» 7.25 » 7.49	12	0,114	99,425
» 7.50 » 7.74	16	0,153	99,578
» 7.75 » 7.99	10	0,095	99,673
» 8 00 » 8 24	4	0,040	99,713
» 8.25 » 8.49	3	0,029	99,742
» 8.50 » 8.74	1	0,009	99,751
» 8.75 » 8 99	2	0,019	99,770
» 9.00 » 9.24	2	0,019	99,789
» 9.25 » 9.49	1	0,009	99,798
» 9 50 » 9 74	1	0,009	99,807
» 9.75 » 9.99	3	0,029	99,836
» 10.00 » 10.24	3	0,029	99,865
» 10.25 » 10 49	1	0,009	99,874
» 10.50 » 10.74	5	0,047	99,921
» 10.75 » 10.99	0	0,000	—
» 11.00 » 11.24	6	0,060	99,981
» 11 25 » 11.49	0	0,000	—
» 11 50 » 11 74	0	0,000	—
» 11 75 » 11.99	1	0,009	99,990

Sur une feuille de papier quadrillé, nous reportons ensuite les mesures suivantes : d'abord nous traçons une ligne horizontale d'une longueur convenable ; cette ligne représente l'ensemble des parts proportionnelles, donc 100.00 ; Galton les appelle « grades ». Cette ligne est divisée de 10 en 10 parties et si l'on peut se servir de papier

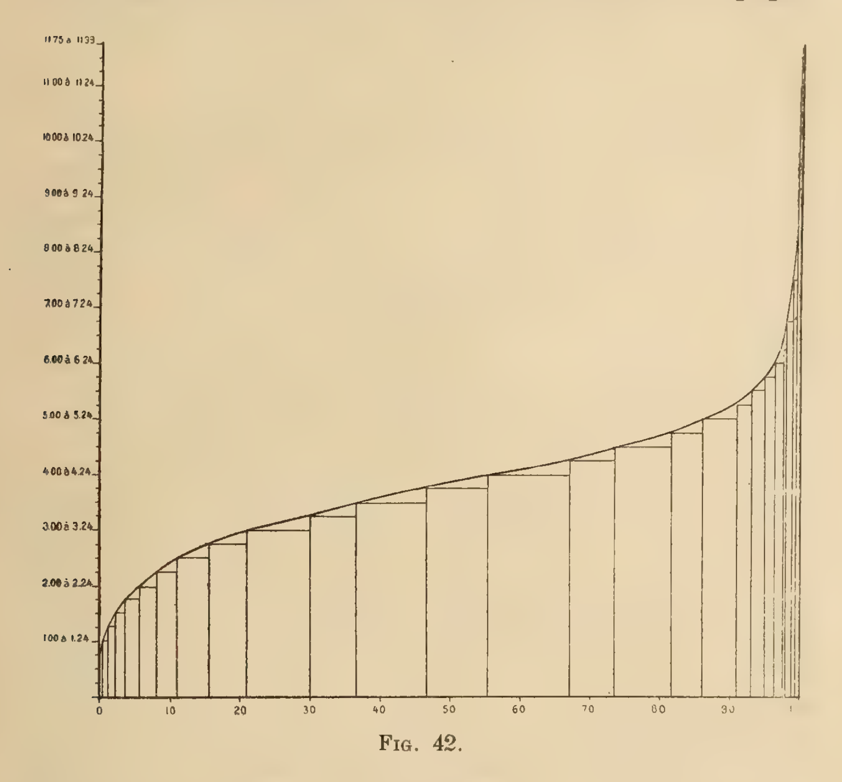


Fig. 42.

millimétrique il est bon d'admettre que 1 mm. ou 2 mm. correspondent à 1 p. c.

On élève ensuite à droite et à gauche une ligne verticale sur laquelle on porte les taux des salaires ; on tiendra compte de la hauteur totale à donner à l'échelle et on divisera cette hauteur en parties égales d'après les variables du taux des salaires.

Revenant à la ligne horizontale, on élèvera sur cette ligne une perpendiculaire à l'endroit correspondant au

chiffre des pourcentages accumulés, d'abord pour la première donnée, puis pour la seconde, etc. La hauteur de ces lignes devra correspondre au taux de salaire en rapport avec la donnée, l'échelle placée à droite ou à gauche guide le dessin à cet égard. Du sommet de ces perpendiculaires on tirera des lignes horizontales de façon à former autant de rectangles. Enfin, les angles de gauche des rectangles seront joints par une courbe partant de l'origine la plus régulière possible.

Pour déterminer le salaire médian, il suffit alors d'élever du point 50, sur la ligne horizontale une perpendiculaire qui s'arrête à la courbe; on lira sur l'échelle des salaires, à la même hauteur, le chiffre du salaire médian; on opère de même pour les quartiles en élevant les perpendiculaires respectivement aux points 25 et 75.

Tout point quelconque peut être connu immédiatement par une simple lecture. De cette façon, on pourra non seulement fixer la position de la médiane et des quartiles, mais encore celle des déciles et des percentiles. La figure a donc une grande puissance représentative.

B. — *Méthode graphique pour la recherche de la médiane, de M. Yule.*

393. Nous avons vu que, dans le calcul de la médiane, il y avait lieu de procéder à une interpolation arithmétique afin de localiser la médiane exactement à l'endroit qui lui convient entre les intervalles de classes. M. Yule a montré que par la construction d'une figure graphique très simple on pourrait rapidement trouver la médiane sans recourir à aucune interpolation (1). Le procédé qu'il emploie à cet effet peut être décrit comme suit : sur une feuille de papier quadrillé, portez en abscisses les divisions de classe et en ordonnées le nombre des fréquences, notez au moyen d'un

(1) G. U. Yule, *An introduction to the theory of statistics*, p. 118.

point l'endroit précis des fréquences comprises entre chaque classe ; réunissez ces points par une ligne courbe à l'aide d'un instrument approprié ; ensuite, sur l'axe des ordonnées marquez un point correspondant au nombre total de fréquences divisé par 2 ($\frac{N}{2}$) ; de ce point, tirez une ligne horizontale par rapport à l'axe des abcisses : le point où la ligne courbe coupe cette ligne horizontale correspond à la médiane, dont vous lirez la valeur sur l'axe des abcisses à l'endroit indiqué par une ligne verticale menant du point de rencontre à l'axe des abcisses. D'après M. Yule, ce procédé donne un résultat tellement précis qu'il dispense le plus souvent de tout calcul arithmétique et notamment du calcul d'interpolation auquel il faut avoir recours d'habitude.

C. — *Méthode graphique pour la recherche du degré de corrélation de Sir Francis Galton.*

394. Comme illustration de l'emploi de la méthode graphique en vue de la recherche d'un résultat statistique, on peut encore citer la méthode graphique de Galton, dans la recherche de la corrélation entre deux phénomènes donnés. Le premier point consiste à déterminer quel est celui des deux phénomènes qui sert de mesure à l'autre : le choix dépend avant tout de la logique ; s'il s'agit d'un exemple emprunté aux sciences sociales on soumettra le cas à une analyse économique serrée. Après avoir arrêté quel est le phénomène principal ou étalon et lequel est le relatif, on cherche de part et d'autre la moyenne et l'on transforme en une série d'index-numbers, calculés par rapport à la moyenne, les données primitives. Les index-numbers étant obtenus, on peut procéder à la recherche graphique du coefficient de covariation. On portera en ordonnée l'échelle des index du fait principal et en abcisses ceux du fait relatif : l'échelle sera graduée de manière à avoir, de part et d'autre, le même intervalle de classe. Toutes les données de

l'index du fait principal seront ensuite dépouillées et localisées sur le diagramme, en tenant compte de la donnée inscrite en regard pour le phénomène relatif. Ainsi la donnée 56, valeur d'un indice de la donnée principale, combinée avec la valeur 97, indice, pour la même année, du phénomène relatif, signifie qu'il faudra porter sur le diagramme un signe distinctif à l'intersection de l'ordonnée 56 avec l'abcisse 97. En général, les points correspondant à l'emplacement des données ont une tendance à se grouper vers la gauche du graphique. Après l'inscription de toutes les valeurs de l'index du phénomène principal, on trace à travers le diagramme une droite inclinée à 45°, qui est appelée ligne d'égale variation; puis on trace une seconde ligne, nommée ligne de régression, de manière à ce qu'elle suive le plus exactement possible la direction générale indiquée par l'emplacement des signes distinctifs sur le diagramme : la direction est exactement donnée lorsque ces signes distinctifs sont groupés le long de la ligne, ou quand la ligne de régression en laisse autant à droite qu'à gauche. Plus les lignes d'égale variation et de régression sont rapprochées, plus le degré de covariation des deux phénomènes est élevé. Pour calculer ce degré de covariation, il faut, d'un point de l'axe des ordonnées, mesurer une droite A B parallèle à l'axe des abcisses, coupant les lignes de régression et d'égale variation. On obtient ainsi une figure triangulaire renversée dont la base est la ligne susdite A B, le grand côté la ligne de régression (A C) et l'autre côté la ligne verticale des ordonnées (B C). On a alors la proportion $\frac{AB}{BC}$, ou la tangente de la figure triangulaire décrite par la droite menée de l'axe des ordonnées, cet axe des ordonnées lui-même et la ligne de régression depuis son origine jusqu'au point d'intersection A.

La valeur de A B et de B C se mesure facilement en tenant compte du nombre de classes et de la valeur de l'intervalle de classe. La fraction $\frac{AB}{BC}$ représente le degré de covariation; si elle est de 0.80 par exemple, elle signifie que

pour tout changement du fait principal, changement équivalant à l'unité, il y a un changement du phénomène relatif équivalant à 80 p. c. du changement type du fait principal. L'équation de régression est représentée par la différence de l'équation de covariation par rapport à l'unité, soit dans le cas imaginé plus haut, 0.20 p. c. (1).

395. — *Références.*

Album de statistique graphique (1900). Paris, Imprimerie Nationale, 1906.

Atlas de statistique financière (1881-1889). Paris, Ministère des Finances.

Atlas statistique du Recensement général des industries et des métiers (1896). Bruxelles, Ministère de l'Industrie et du Travail, 1901.

Auerbach (F.), « Die graphische Darstellung », aus *Natur und Geisteswelt*, Nr. 437, Leipzig, Teubner.

Benini (R.), *I diagrammi a scala logaritmica,* in « Festgabe für Adolf Wagner », Leipzig, 1905.

Benini, « Principii di statistica metodologica », Torino, 1906, pp. 133-156.

Bertillon et autres, « Uniformité dans l'établissement des graphiques », *Bulletin de l'Institut international de statistique*, vol. XIII (session de Budapest, 1901).

Bosco (A.), « Lezioni di Statistica », Roma, 1909.

Brinton (W.-C.), « Graphic methods for presenting facts », *Engineering Magazine* Co, 1914, New-York.

Cheysson (E.), « Histoire d'un tableau statistique », *Revue scientifique*, 1888.

Cheysson (E.), « La statistique géométrique », *Œuvres choisies*, Paris, 1911, t. I, pp. 185-218.

Field (J.-A.), « Some advantages of the logarithmic scale in statistical diagrams », *Journal of political Economy*, octobre 1917.

Fisher, « The ratio chart », Quarterly Publi. of the amer. Statist. Association, 1917, p. 577.

Isserlis, « On the representation of statistical data », *Biometrika*, mai 1917, pp. 418-425.

Kowatsch, « Illustrirte deutsche Statistik, Diagramme und Stufenkarten », Berlin, Puttkammer und Muehlbrecht, 1912.

---

(1) On trouvera un exemple de cette méthode, accompagné d'un graphique, dans le manuel de King, *Statistical methods.*

Majorana-Calatabiano, « Teoria della statistica », Roma, 1889, pp. 180 et suiv.

March (L.) et autres, « Les moyens de rendre comparables, les courbes statistiques », *Bulletin de l'Institut international de statistique*, vol. XIX (session de La Haye, 1911).

Mayr (G. von), « Statistik und gesellschaftslehre, *Theoretische statistik* », Freiburg, 1895, p. 102 et suiv.

Mayr (G. von), « Zur Methodik und Technik statischer Karten », *Allgemeines Statistisches Archiv.*, vol. 7, 1914, pp. 131-157.

Minguez y Vincente, « Tratado de estadistica », Cordoba 1898.

Peddle (J.-B.), « The construction of graphical charts », Mac Graw Hill, 1910, New-York.

Perozzo, « Della reppresentazione grafica di una collettivita di individui nella successione del tempo, e in particolare dei diagrammi a tre coordinati », *Annali di Statistica*, Roma, 1880.

Perozzo, « Stereogrammi demografici », *Annali di Statistica*, Roma, 1881.

Pirani (M. von), « Graphische Darstellung in Wirtschaft und Technik », Berlin, Göschen, 1914.

Roesle, « Graphisch-statistische Darstellungen, *Beilage zum deutschen Statistischen Zentralblatt*, 1913, Nr. 2, pp. 25-39.

Schmidt (G.-H.), « Kartographische Darstellung der Volksdichtigkeit », *Allgemeines statistisches Archiv.*, vol. 7, 1914 pp. 158-163.

Schott (S.), « Graphische Darstellungen », *Die statistik in Deutschland*, München, Sellier, 194, t. I, pp. 187-194.

Secrist (H.), « An introduction to statistical methods, Macmillan, 1917, pp. 158-233 (avec références), New-York.

Whipple (G.-C.), « Vital statistics », Wiley, 1920, pp. 58-99, New-York.

LIVRE III

LA LOI DES ERREURS

I. — Définitions et généralités.

396. Dans le langage mathématique, on donne le nom d'erreur à la différence entre une quantité, d'ordinaire inconnue, mais tenue pour exacte, et l'évaluation qu'on en fait. L'erreur mathématique est donc le résultat d'une observation mal faite. On distingue l'erreur absolue qui, par excès ou défaut, est commise par rapport à un nombre et est considérée en elle-même, sans avoir égard à ce nombre; on distingue encore l'erreur relative, qui exprime le rapport de l'erreur elle-même au nombre d'après lequel elle est estimée; l'erreur, divisée par le nombre exact, a pour résultat l'erreur relative. La théorie des erreurs forme une partie intéressante des mathématiques et elle est à la base de la théorie des probabilités. Le langage philosophique donne au mot erreur un sens différent de celui qui vient d'être exposé : l'erreur consiste à affirmer comme vraie une chose qui n'est pas, ou à nier une chose qui est. L'erreur philosophique n'est pas sujette à mesure; elle ne comporte pas une série de degrés susceptibles, dans leur grandeur et leur arrangement, d'obéir à une certaine loi; par là, elle diffère essentiellement de ce qu'on appelle « erreur » dans les sciences mathématiques. La statistique donne un sens spécial aux mots « erreur » et « loi des erreurs », — nous verrons plus loin en quoi ce sens est spécial; notons

cependant que le sens attribué à ces mots par la statistique se rapproche très sensiblement de celui du terme « erreur » en termes mathématiques.

La « loi des erreurs » est une formule qui désigne en abrégé certaines propriétés des erreurs, notamment celle d'obéir à des groupements systématiques.

Elle porte plusieurs noms, sous lesquels elle est indifféremment désignée : loi des grands nombres, loi de Gauss, courbe de probabilité, etc. Nous conserverons au cours de cet exposé, la dénomination de « loi des erreurs » — parce qu'elle est la plus compréhensive et la plus exacte de celles proposées.

397. Avant d'aborder l'examen du fond, il nous faut expliquer le sens statistique du mot « erreur » et justifier l'emploi de l'expression : « loi des erreurs ».

« Aucune observation, écrit Bertrand, n'est certaine, mille opérations successives donnent mille résultats différents. Non que l'observateur, de mieux en mieux instruit, corrige ses défauts et s'avance vers la perfection. Il n'en est pas ainsi. Les derniers résultats ne ressemblent en rien à une limite dont on s'approcherait par continuel progrès; les évaluations, tantôt trop petites, tantôt trop grandes, se succèdent en confusion et sans ordre comme des boules blanches ou noires puisées dans une urne. »

Bertrand, dans ce passage, attire l'attention sur un fait remarquable, à savoir que les erreurs se succèdent sans ordre; il ne s'agit pas d'une suite d'erreurs qui, considérables au début à cause de l'inexpérience de l'opérateur, iraient en diminuant au fur et à mesure que l'habileté de l'observateur irait en augmentant; il semble que le hasard seul règle la fréquence et la grandeur des erreurs. S'il existait, parmi les causes d'erreur, des motifs permanents de se tromper, les erreurs resteraient constantes et se fixe-

raient dans le même sens ; mais à côté de ces erreurs, on conçoit parfaitement qu'il puisse en exister qui ne se présentent qu'occasionnellement. De là la division des erreurs en erreurs systématiques et erreurs accidentelles (1).

Les erreurs systématiques sont celles qui sont sujettes à répétition dans un sens connu ; elles sont personnelles si elles tiennent à la personne même de l'observateur, instrumentales si elles sont dues à une particularité de l'instrument employé dans l'observation, théoriques si elles dérivent de certaines causes dont les conséquences peuvent se peser à l'avance. Les erreurs personnelles sont la conséquence de ce qu'on appelle « l'équation personnelle » de l'opérateur ; les erreurs instrumentales se retrouvent dans les mesures physiques où l'on fait usage d'instruments compliqués et délicats ; en statistique, ces deux catégories d'erreurs n'ont guère l'occasion de se produire ; cependant, on pourrait assimiler à une erreur instrumentale celle qui résulte, dans un travail d'investigation statistique, d'une question mal posée, d'instructions incomplètes ou peu claires, d'une méthode vicieuse ; il est vrai qu'on peut aussi, pour de bonnes raisons, les ranger parmi les erreurs théoriques (2).

Disons immédiatement que ce n'est pas des erreurs systématiques, quelle que soit leur origine, qu'il s'agit dans la « loi des erreurs ». Elles sont exclues par définition du calcul des probabilités, lequel ne s'applique qu'aux seules erreurs accidentelles.

On entend par erreurs accidentelles, celles qui se présentent encore dans les observations les plus soigneusement purgées de toute erreur systématique. « Les erreurs accidentelles, dit Gruey, ne sont liées par aucune loi aux

---

(1) On dit aussi « erreurs régulières » au lieu de « erreurs systématiques ».

(2) Il est superflu de faire remarquer qu'une série de mesures affectée d'erreurs régulières ou systématiques doit être immédiatement rejetée.

observations et ne peuvent être calculées *a priori*. N'étant pas calculables *a priori* à cause de la nature capricieuse et irrégulière des causes qui les produisent, elles tombent dans le domaine du calcul des probabilités (1). »

Le fait que les erreurs accidentelles sont soumises aux règles du calcul des probabilités résulte du théorème de Bernouilli dont nous parlerons plus loin ; mais auparavant, il est permis d'invoquer l'expérience et il n'est pas inutile, pour préciser les idées, d'exposer à l'intention du lecteur quelques faits à l'appui de l'affirmation qui précède.

L'astronome anglais James Bradley (1693-1762) ayant mesuré 470 fois, avec toute la précision possible, le même intervalle de temps, il se trouva que, malgré le soin avec lequel elles avaient été exécutées, ces observations étaient encore entachées d'erreurs, les unes ayant dépassé le temps moyen, les autres étant restées en deçà : ainsi, lorsque Bessel en fit le classement, il trouva que :

94	observations	présentaient	une	erreur	comprise entre	0.0 et 0.1
88	»	»	»	»	»	0.1 et 0.2
78	»	»	»	»	»	0.2 et 0.3

et ainsi de suite jusqu'à une erreur de 0.9 à 1.0, le nombre des erreurs allant en diminuant au fur et à mesure que leur importance augmentait.

(1) Cette observation est exacte au point de vue subjectif, mais il semble que l'analyse est poussée plus loin par M. Mansion lorsqu'il écrit : « les erreurs irrégulières, accidentelles ou fortuites considérées *objectivement* se reproduisent identiquement dans les mêmes circonstances, absolument comme les erreurs régulières ; autrement dit, en soi, toutes les erreurs sont régulières. Mais *subjectivement*, par rapport à nous, il n'en est pas de même : lorsque nous croyons que les circonstances sont les mêmes, elles ont changé, mais trop peu pour que nous puissions le savoir ; il y a de petites causes d'erreur, agissant avec une égale facilité dans tous les sens, qui introduisent des erreurs que nous ne pouvons pas constater. Ce sont les erreurs que l'on appelle irrégulières, accidentelles ou fortuites. » BOUDIN-MANSION, *loc. cit.* p. 154.

Un exemple à peu près analogue a été donné par Quetelet : utilisant plusieurs centaines d'observations faites à l'Observatoire de Greenwich, entre 1836 et 1839, Quetelet a trouvé que les résultats de l'observation s'écartent de la moyenne à raison de certaines circonstances faisant naître des erreurs accidentelles. Les observations dont Quetelet a fait usage sont tirées des publications de l'Observatoire royal de Greenwich et concernent les déterminations en temps de l'ascension droite de la polaire, ces mots ascension droite signifiant la distance de l'astre au point équinoxial, mesurée le long de l'équateur céleste; ces observations, comme le fait remarquer Quetelet, ont subi différentes corrections pour la nutation, la précession, etc., et ont été calculées pour une même époque, en sorte qu'elles ne diffèrent que par les effets des petites erreurs accidentelles (1).

Le tableau dressé par Quetelet montre que sur 487 observations il se trouve :

145	observations	présentant	une	erreur	comprise entre	0.0 et 0.5
124	»	»	»		»	0.5 et 1.0
74	»	»	»		»	1.0 et 1.5
37	»	»	»		»	1.5 et 2.0
11	»	»	»		»	2.0 et 2.5
2	»	»	»		»	2.5 et 3

82 observations correspondaient exactement à la moyenne et ne présentaient par conséquent aucune erreur appréciable.

On voit, d'après ce qui précède, que la détermination de l'importance et du sens des erreurs dans les mesures expérimentales se fait en comparant chaque résultat à la

(1) QUETELET : *Lettres sur la théorie des probabilités*, pp. 125-126. Bruxelles, 1846.

moyenne calculée sur l'ensemble des observations. Ainsi, pour 470 observations faites par Bradley pour la déclinaison d'une étoile, la valeur fixe adoptée pour la déclinaison est la moyenne de toutes les observations faites; les écarts, c'est-à-dire les différences entre cette moyenne et chaque résultat particulier ont été calculés avec une grande précision, puis ils ont été classés suivant une échelle dont chaque division correspond à 0″,4 (1). Le même procédé a été appliqué par Quetelet aux calculs de l'ascension droite de la polaire, mais avec une plus large approximation.

398. Pourquoi admet-on que la moyenne puisse servir à déterminer l'erreur de toute observation particulière? C'est parce que, comme nous l'avons démontré plus haut, entre toutes les mesures prises d'un phénomène quelconque, la moyenne constitue la valeur la plus exacte. Le lecteur devra se référer à l'application donnée antérieurement, mais pour encore préciser ses idées à ce sujet, nous reprendrons une explication sommaire que nous empruntons à M. Thomas Wallace Wright et qui a le grand mérite de la simplicité (2).

Supposons que x soit la quantité à mesurer et que d soit le résultat de l'observation.

S'il n'y a aucune erreur dans l'observation faite, on a

$$x - d = 0$$

Mais si, après avoir fait plusieurs observations de cette même quantité, nous trouvons que les mesures obtenues diffèrent entre elles, nous devons conclure nécessairement que plusieurs d'entre elles sont inexactes; désignons par la

---

(1) E. CARVALLO : *Le calcul des probabilités et ses applications*, p. 90, Paris-Gauthier-Villars, 1912.

(2) TH. WALLACE WRIGHT : *The adjustment of observations*, 2e édit., London et New-York, 1906, p. 9 et suiv.

lettre Δ les erreurs d'observation et affectons chaque résultat d'observation et chaque écart ou erreur d'un indice permettant de les reconnaître ; nous avons alors :

$$\begin{aligned} x - d_1 &= \Delta_1 \\ x - d_2 &= \Delta_2 \\ x - d_3 &= \Delta_3 \\ x - d_n &= \Delta_n \end{aligned} \qquad (81)$$

La valeur à chercher est la véritable signification de d et nous devons admettre que cette valeur doit être considérée comme une fonction des résultats fournis par l'observation : elle est comprise entre la valeur la plus élevée et la valeur la moins élevée de d. On peut admettre que si ces valeurs étaient rangées par ordre d'importance, la valeur idéale se trouverait vers le centre, au point également éloigné des deux extrêmes, c'est-à-dire à la valeur centrale si le nombre des observations est impair, et aux deux valeurs centrales si le nombre d'observations est pair ; le lecteur voudra bien se rappeler ce que nous avons dit plus haut au sujet de la médiane. Bien entendu, nous devons supposer toujours que toutes les observations sont également dignes de foi et qu'aucune d'entre elles n'est entachée d'une erreur systématique ; nous sommes et nous devons rester dans le domaine des erreurs accidentelles. Or, la fonction symétrique la plus simple qui puisse être choisie pour x_0 représentant la fonction symétrique de x, est la moyenne arithmétique

$$x_0 = \frac{(d_1 + d_2 + d_3 \ldots + d_n)}{n} = \frac{\Sigma\,(d)}{n} \qquad (82)$$

d'où il suit que, entre toutes les valeurs fournies par l'observation, la plus exacte est celle qui est fournie par la moyenne arithmétique des résultats livrés par l'observation.

399. Nous nous sommes placé jusqu'à présent dans l'hypothèse de mesures idéales ayant rapport à des nombres déterminés par l'analyse mathématique, ou à des mesures réelles, expérimentales, ayant trait à des phénomènes naturels. Mais nous sommes, ne l'oublions pas, dans le domaine des faits concrets, dans le domaine de la statistique et une question importante se pose à l'instant : de quel droit transpose-t-on dans ce domaine de la statistique les formules et les théorèmes reconnus exacts dans la sphère du calcul pur ou dans celle des observations relatives aux phénomènes naturels? La question nous semble intéressante et vaut qu'on s'y arrête un instant. Déjà Jacques Bernouilli avait aperçu que le calcul des probabilités s'étendait à d'autres matières que celle des jeux de hasard : la 4e partie de son « Ars conjectandi » a en effet pour titre: « De l'emploi et de l'application de la doctrine précédente dans les matières civiles, morales et économiques », mais la mort n'a pas permis au grand mathématicien d'énoncer ses vues au sujet de l'application des probabilités au vaste domaine qu'il assignait à sa « doctrine ». Un fait d'expérience, d'une portée considérable, est qu'un grand nombre de phénomènes naturels ont une distribution autour de la moyenne telle que leurs éléments semblent être rassemblés par l'effet du hasard; on pourrait presque les assimiler aux boules blanches ou noires tirées hors d'une urne, dont la loi de sortie est déterminée par le calcul des probabilités. Citons une fois encore les chiffres relatifs à la taille de 8,585 adultes mâles nés en Angleterre, Irlande, Ecosse et Pays de Galles, d'après le rapport final de la commission d'anthropométrie adressé à la British Association (rapport de 1883, cité par M. G. U. Yule, p. 88) :

HAUTEUR SANS BOTTINES — Pouces	NOMBRE DES HOMMES COMPRIS DANS LES LIMITES CI-CONTRE : ENDROIT DE LA NAISSANCE				
	Angleterre	Ecosse	Galles	Irlande	Total
57	1	—	1	—	2
58	3	1	—	—	4
59	12	—	1	1	14
60	39	2	-	—	41
61	70	2	9	2	83
62	128	9	30	2	169
63	320	19	48	7	394
64	524	47	83	15	669
65	740	109	108	33	990
66	881	139	145	58	1223
67	918	210	128	73	1329
68	886	210	72	62	1230
69	753	218	52	40	1063
70	473	115	33	25	646
71	254	102	21	15	392
72	117	69	6	10	202
73	48	26	2	3	79
74	16	15	1	—	32
75	9	6	1	—	16
76	1	4	—	—	5
77	1	1	—	—	2
Totaux	6194	1304	741	346	8585

La distribution par grandeur de taille n'est pas identique à celle d'une distribution symétrique idéale, mais il faut reconnaître qu'elle s'en rapproche singulièrement. La moyenne est localisée entre les grandeurs 67 et 68 pouces, exactement 67,46 correspondant à la fréquence la plus élevée : 1,329 individus. Examinant un exemple analogue, M. Borel suppose que l'on mesure la taille de tous

les Français adultes : « Si l'on convient, dit-il, de considérer la valeur moyenne des mesures comme la valeur exacte que devrait avoir la taille d'un Français, on constate que les « erreurs », c'est-à-dire les différences positives ou négatives entre cette valeur théorique et la valeur réelle se répartissent précisément suivant la loi de Gauss : tout se passe comme si les Français avaient tous la même taille, égale à la moyenne, mais étaient mesurés par un expérimentateur très maladroit, dont les erreurs de mesure suivraient la loi de Gauss (1). »

Des exemples de ce genre peuvent être relevés en grand nombre parmi les mesures anthropométriques. On en connaît aussi dans le domaine des sciences zoologiques et botaniques; en admettant une certaine latitude d'appréciation, on en trouve aussi dans la sphère des phénomènes sociaux : il ne faut pas penser que la courbe des salaires construite en tenant compte du nombre des ouvriers qui gagnent un salaire d'un certain taux, soit absolument irrégulière; si elle n'épouse pas absolument la courbe symétrique idéale, elle s'en rapproche dans une mesure assez notable pour que cette approximation soit de nature à intéresser vivement les esprits réfléchis.

Mais il s'agit de ressemblance, de rapprochement et non de similitude. Les phénomènes humains, a dit quelque part Borel, ressemblent à la loi des erreurs, à peu près comme une orange ressemble à une figure géométrique idéale, la sphère. Et la ressemblance subit une sorte de dégradation à mesure qu'on s'élève dans l'échelle des êtres ou qu'il s'agit de manifestations de plus en plus conscientes. Il ne peut donc être question d'expliquer par la loi des erreurs tous les phénomènes qui se présentent d'après un plan plus ou moins analogue; il est seulement légitime, pour certains d'entre eux, de les rapprocher des types géométriques de

(1) Emile Borel, *Eléments de la théorie des probabilités*, p. 184, Paris, A. Herman et Fils, 1909.

distribution, et pour d'autres d'aller plus loin et de calculer, d'après les méthodes qui seront indiquées ci-après le degré de probabilité qu'ils présentent. Les phénomènes sociaux, particulièrement, ne peuvent être interprétés qu'avec beaucoup de précaution. On sait que Quetelet imaginait une assimilation systématique des faits naturels à la loi des erreurs. L'illustre statisticien était tenté de voir dans la nature une urne dont la composition est inconnue; blanches ou noires, les boules tirées apparaissent dans un désordre apparent, et, en effet, les événements de la vie, par leur bizarrerie et leur caractère inattendu, semblent échapper à toute prévision. Mais la loi des erreurs rend compte des raisons de leur sortie, et après qu'ils ont été convenablement classés, l'ordre le plus parfait règne parmi des événements qui ne paraissent obéir à aucune tendance systématique. Ce système philosophique a pour lui un certain nombre de faits impressionnants, mais il en a tout autant contre lui. Ce qu'il importe de retenir, c'est que la loi des erreurs est d'une importance capitale. mais qu'on ne saurait, sans tomber dans l'exagération, l'étendre à tous les phénomènes et systématiser son empire. Pour qu'elle s'applique aux phénomènes, ceux-ci doivent présenter des variations dues à des circonstances accidentelles et dès lors soumises au calcul des probabilités; cette condition exclut certains faits dans lesquels intervient la volonté consciente de l'homme et doit nous rendre prudents dans l'examen des faits économiques.

II. — **Notions sur les probabilités.**

Théorème de Bernouilli. — Règle des écarts.

400. Si nous supposons qu'un événement peut se produire dans x cas et que tous ces cas sont également possibles, nous nous trouvons dans les conditions voulues pour définir la probabilité: la probabilité d'un événement donné

s'exprime par le rapport des cas favorables à cet événement au nombre de cas possibles.

On vient de voir que la définition exige, pour le calcul, la connaissance de deux points importants : le nombre de cas favorables et le nombre de cas possibles. Pour déterminer le nombre de cas également possibles, on aura recours au calcul des combinaisons dont nous avons parlé dans l'Introduction à ce volume ; ce calcul est souvent laborieux et la manière de poser le problème soulève maintes fois de sérieuses difficultés. Le calcul des probabilités est ardu, non pas tant par ses difficultés techniques, qui sont, après tout, une affaire de formule, que par la difficulté logique de poser le problème d'une façon exacte. La possibilité égale des événements suppose qu'on procède à de longs calculs : ainsi, en admettant qu'on joue à un jeu n'admettant que deux alternatives, tel que le jeu de pile ou face, et en supposant qu'on se livre à 100 épreuves, on devra, pour obtenir le nombre de cas également possibles, calculer le produit de 100 multiplicateurs égaux à 2 ($\overset{(1)}{2} \times \overset{(2)}{2} \times \overset{(3)}{2} \times \overset{(4)}{2} \times \overset{(5)}{2} = 32 \ldots \times \overset{(100)}{2}$), produit qu'on exprime par $\overset{(100)}{2}$. On éprouve les mêmes difficultés de calcul en ce qui regarde l'énumération des cas favorables ; les combinaisons qu'on peut obtenir en choisissant x cas parmi 100 ou 200 objets donnés, sont excessivement nombreuses ; nous avons dit plus haut que la formule algébrique qui s'applique à ces cas est $P — 1.2.3\ldots n$, c'est-à-dire que chaque produit est multiplié par le nombre suivant, jusqu'à ce qu'on soit arrivé au chiffre n ; si $n = 21$, on aura $1 \times 2 = 2$; $2 \times 3 = 6$; $6 \times 4 = 24$; ... $n \times 21 = x$.

401. La probabilité qui vient d'être définie s'appelle probabilité *a priori* parce qu'on suppose que ses éléments sont connaissables, sans autre recherche que celle des calculs ; dans le jeu de pile ou face, par exemple, on sait qu'il n'y a d'autre alternative que d'obtenir pile, ou d'obtenir face ;

au jeu de dé, comme il y a six faces portant chacune un numéro, on sait d'avance que le cas favorable à la sortie de l'un de ces numéros est de 1/6, en supposant, bien entendu, que le dé ne soit pas faussé.

Mais il est une autre espèce de probabilité qu'on appelle probabilité *a posteriori* et à laquelle on pourrait aussi donner le nom de probabilité statistique, tant elle est fréquente dans le domaine de cette méthode scientifique. Si, dans certaines hypothèses, empruntées au jeu le plus souvent, on peut, à l'avance, faire connaître les conditions de la probabilité, il n'en est pas de même dans la plupart des événements de la vie.

Pour connaître la probabilité qu'une naissance soit masculine, plutôt que féminine, il a fallu évidemment noter le sexe de l'enfant dans un grand nombre de naissances; pour savoir quelle est la chance qu'une maison brûle, qu'un navire périsse, qu'un accident se produise, il a été nécessaire d'observer un grand nombre de sinistres et de comparer leur fréquence à la totalité des cas existants. Cette recherche est faite par la statistique, et la probabilité qui résulte de recherches de la statistique s'appelle probabilité *a posteriori*. Sans la statistique, on ne saurait mesurer la probabilité des événements soumis à des influences complexes parce que le nombre de cas favorables à l'événement ne peut alors être connu que par l'observation.

Ici, encore, il convient d'attirer l'attention sur la différence essentielle entre la probabilité statistique et la probabilité *a priori* : la probabilité statistique ne nous donne qu'une notion approximative qui n'est nullement comparable aux probabilités abstraites et rigoureuses auxquelles on arrive en tablant sur les résultats, par exemple, du jeu de pile ou face ou sur la loi de sortie des boules blanches ou noires d'une urne d'une composition connue. La probabilité statistique n'est qu'une approximation à l'égard de la seconde.

402. On vient de voir que la probabilité s'exprime par le rapport des cas favorables aux cas possibles. Mais doit-on conclure de là qu'ayant calculé la probabilité sur un grand nombre de cas, cette même probabilité se vérifiera pour chacune des séries qui composent ensemble la totalité des cas possibles? Le bon sens se refuse à l'admettre; on sait très bien que la probabilité est la résultante d'une foule de circonstances accidentelles et que si elle est vraie pour l'ensemble, elle ne l'est certes pas pour chacune des épreuves ou des séries dont se compose le total. On ne comprendrait pas que la probabilité devînt une fatalité. Dans une série de 100 épreuves au jeu de pile ou face, la fréquence 0, c'est-à-dire l'expression fractionnaire obtenue en divisant le nombre des apparitions de l'événement attendu par le nombre de cas possibles (0/100) ne se présentera pas, car il est absolument invraisemblable que sur 100 épreuves on réussisse à obtenir 100 fois pile ou 100 fois face, selon que l'on aura parié pour pile ou pour face; au contraire la fréquence 50/100 qui représente une égale probabilité pour pile et pour face a pour elle une énorme probabilité.

En procédant par voie expérimentale nous obtiendrons donc des résultats différents pour chacune des séries par rapport à leur fréquence; les unes se rapprocheraient d'une manière notable de la probabilité théorique générale, les autres s'en éloigneraient beaucoup. La probabilité et la fréquence, fait observer Carvallo dans son remarquable exposé (1), sont deux termes qui ont des natures presque opposées. La probabilité a un caractère spéculatif et son existence objective est toujours douteuse, parce qu'on n'est jamais certain que les divers cas envisagés sont également possibles; mais supposez-vous qu'ils soient tous possibles au même degré, parce que, par exemple, l'opération de statistique, l'observation scientifique, a été

(1) CARVALLO : *Le calcul des probabilités et ses applications*, Paris, Gauthier-Villars, 1912, p. 8.

faite de bonne foi, par un homme capable, au moyen d'une bonne méthode et avec des instruments convenables, il faut admettre que la probabilité est certaine et qu'à mesure que se multiplieront les essais, les épreuves, le résultat se rapprochera davantage de cette probabilité. Au contraire, la fréquence a une valeur objective, mais cette valeur est incertaine, elle est variable, elle change dans une multitude de cas : tantôt elle se rapproche de la probabilité, tantôt elle s'en éloigne.

De là, on tire une nouvelle notion, celle de l'écart. L'écart d'une série d'épreuves est la différence entre la probabilité théorique de l'événement attendu et la fréquence de l'apparition du même événement dans une série donnée d'observations (1).

De même, à propos de la moyenne, on appelle écart, déviation, la différence numérique entre tout nombre de la série et la moyenne. Ici, la probabilité est la moyenne, la fréquence est la valeur attribuée par l'observation dans chaque cas particulier, et l'écart, dans un cas comme dans l'autre, est la différence entre ces deux termes.

403. Maintenant essayons d'introduire une nouvelle notion. Ces écarts, — qu'on les considère à propos de la moyenne ou de la probabilité, peu importe, car leur nature, dans les deux cas, est identique —, ces écarts ont-ils la chance de se grouper d'une manière systématique autour de la valeur à laquelle ils se comparent? La logique semble l'indiquer, car on se refuse à croire que dans une multitude d'épreuves, on atteigne toujours le même résultat, c'est-à-dire que l'écart soit permanent, d'une fixité complète. Chacun s'attend, au contraire, à ce que les écarts soient différents, mais on pourrait croire qu'ils se présentent en désordre. Pour vérifier cette supposition, l'expérience peut

(1) Carvallo : *Le calcul des probabilités et ses applications*, Paris, Gauthier-Villars, 1912, p. 8.

suffire. Quetelet a eu la patience de procéder à cette expérience. Vingt boules blanches et vingt boules noires ayant été placées dans une urne, Quetelet procéda à une série de tirages, la boule tirée étant chaque fois remise dans l'urne, de manière que toutes les circonstances de l'expérience restassent les mêmes. Les résultats des tirages sont résumés dans le tableau suivant :

NOMBRE DE BOULES TIRÉES	DEGRÉ DE PRÉCISION	NOMBRE DE BOULES		RAPPORT DES NOMBRES PRÉCÉDENTS
		blanches	noires	
4	2	1	3	0.33
16	4	8	8	1.00
64	8	28	36	0.78
256	16	125	131	0.95
1024	32	528	496	1.06
4096	64	2066	2030	1.02

Les chances étant égales, le nombre de boules tirées devrait être égal de part et d'autre, mais il n'en a été ainsi qu'une seule fois et cet accord peut être considéré comme accidentel; au contraire, il y a évidemment tendance à se rapprocher de l'unité par la multiplication des tirages. L'expérience est intéressante, mais on ne peut songer à l'étendre à tous les domaines, ni à la prolonger au delà d'un certain nombre : au sixième tirage, Quetelet a déjà dû perdre beaucoup de temps à retirer 4,096 fois une boule hors de l'urne et à la compter du côté des blanches ou des noires. A la septième épreuve, il aurait pu retirer 16,384 boules, à la huitième 65,536, à la neuvième 262,144 et à la dixième 1,048,576, ce qui met déjà l'expérience prolongée au rang des impossibilités.

Il ne faut donc pas s'attendre à jeter quelque lumière sur la question au moyen de l'expérience : celle-ci ne peut ser-

vir qu'à indiquer la voie, mais la solution appartient à la théorie. Celle-ci a été résumée en une formule célèbre connue sous le nom de théorème de Jacques Bernouilli : étant donné un nombre ϵ aussi petit que l'on veut, la probabilité pour que la différence entre le rapport observé du nombre des événements favorables et le nombre des événements contraires, d'une part, et le rapport théorique $\frac{p}{q}$, d'autre part, soit supérieure en valeur absolue à ϵ, tend vers zéro lorsque le nombre n des épreuves augmente indéfiniment.

L'exposé du théorème est reproduit ici d'après Borel (1) ; on en trouvera un exposé plus complet dans Cournot (2), mais l'énoncé raccourci qui vient d'être donné suffit à faire apparaître les caractères essentiels de la proposition :

1° Si le rapport des événements A aux événements B est égal à la probabilité théorique, cette répartition est la plus probable qui puisse se réaliser; si le rapport s'éloigne de cette probabilité, la série affectée par ce rapport offre une répartition d'autant moins probable que la différence constatée est grande.

2° Si le nombre des épreuves augmente beaucoup, la probabilité pour que la différence entre le rapport observé et le rapport théorique soit supérieur à un nombre ϵ supposé aussi petit que l'on veut, tend vers zéro (3).

---

(1) BOREL : *Eléments de la théorie des probabilités*, Paris, Hermann, 1919, p. 64.

(2) COURNOT : *Exposition de la théorie des chances*, p. 54. M. Mansion, *loc. cit.*, p. 46, donne cet énoncé abrégé du théorème de Jacques Bernouilli : « Soient deux évènements contraires A et B, de probabilités p et q, soumis à m épreuves répétées dans les mêmes circonstances, c'est-à-dire que l'on suppose p et q constants, m le nombre de répétitions de l'évènement A :

[*Pour* μ *suffisamment grand, on a, presque certainement,* $m : \mu =$ *à peu près* p.]

Cfr. la brillante démonstration complète du théorème de Bernouilli donnée par M. Mansion, *loc. cit.*, pp. 47 à 58.

(3) DE MONTESSUS : *Leçons élémentaires sur le calcul des probabilités*, Paris, Gauthier-Villars, 1910.

404. La démonstration du théorème de Bernouilli appartient au calcul des probabilités ; le statisticien n'a qu'à se référer aux traités sur la matière, il en existe d'excellents et parmi les plus récents parus en langue française il en est dont les auteurs ont fait un louable effort pour se mettre à la portée des personnes non familières avec les mathématiques supérieures ou qui n'en font pas un usage habituel.

Le lecteur voudra bien consulter ces ouvrages s'il désire trouver la démonstration du fameux théorème de Bernouilli. Sans empiéter sur un domaine qui n'est pas le nôtre, nous voudrions cependant donner au lecteur une certaine idée du maniement de la formule qu'il faut employer pour déterminer les éléments numériques de la probabilité. Nous avons déjà dit que ces calculs étaient excessivement longs.

Supposons qu'on ait, parmi 100 objets donnés à choisir de toutes les façons imaginables, d'une part 40 et d'autre part 60 objets. Nous sommes dans la règle des combinaisons et nous pouvons écrire comme première expression destinée à calculer le nombre de cas qui feront apparaître l'événement souhaité 40 fois sur 100 :

$$C_{100}^{40} = \frac{P_{100}}{P_{40} \times P_{60}} \tag{83}$$

Voulons-nous savoir en combien de cas l'événement apparaîtra 41 fois, nous aurons, en modifiant en conséquence la formule précédente :

$$C_{100}^{41} = \frac{P_{100}}{P_{41} \times P_{59}}$$

et pour épuiser la série, il nous faudra aller jusqu'à l'inversion des termes P_{40} et P_{60}. L'exemple précédent est basé sur une alternative, comme au jeu de pile ou face, de sorte que le dénominateur de la fraction est représenté par 2_{100}

et le numérateur par la somme de toutes les combinaisons possibles depuis 40 jusqu'à 60 (1) :

$$P_{60}^{40} \frac{C_{100}^{40} + C_{100}^{41} + \ldots + C_{100}^{60}}{2^{100}} \tag{83}$$

Même pour un exemple simple qui ne comprend que les 100 premiers nombres, le calcul des expressions ci-dessus comporte un travail énorme, capable de faire reculer le calculateur le plus intrépide. Il ne faut même pas compter sur le secours des machines à calculer, parce que la grandeur des nombres dépasse rapidement la capacité d'enregistrement de ces appareils; et l'aide des tables de calcul n'est non plus que d'une médiocre utilité à cause que les nombres deviennent en peu de temps trop considérables; il y a bien les procédés de décomposition des nombres, mais ces méthodes compliquées et longues ne peuvent réellement rendre pratiques des calculs aussi étendus.

405. Il importait de trouver une simplification de ces expressions, sinon le calcul des probabilités, sans application possible, fût resté à l'état de simple curiosité. La formule simplificative est due à Stirling, mathématicien écossais (1692-1770); elle a pour but de faire connaître, d'une manière approximative, le produit des n premiers nombres entiers, quel que soit le nombre en question.

La formule de Stirling peut s'écrire de la sorte (2) :

$$n! = n^n e^{-n} \sqrt{2 \pi n}\,(1 + \epsilon n) \tag{84}$$

ou d'une manière plus simple et en désignant le nombre obtenu par Sn (S étant l'initiale de Stirling) (3) :

$$Sn = \left(\frac{n}{e}\right)^n \sqrt{2 \pi n} \tag{85}$$

(1) Ces expressions sont empruntées à M. Carvallo, *op. cit.*, pp. 14-15 dont nous suivons plus d'une fois le remarquable exposé.

(2) Borel, *op. cit.*, p. 39.

(3) Carvallo, *op. cit.*, p. 16. (Valeur approchée seulement tant que n est fini.)

Dans la première formule, dont la démonstration revient aux ouvrages d'analyse, n est un nombre entier aussi grand que de besoin, e est la base des logarithmes nepériens (2,71828...), π est le rapport bien connu de la circonférence au diamètre, ($\pi = 3.14159...$) et ϵn un nombre variable avec n, mais qui tend vers zéro lorsque n augmente indéfiniment (1).

Ces mêmes éléments définissent tous les termes de la seconde formule.

La formule de Stirling est une formule d'approximation grossière. La différence entre le nombre $n!$ donné par la formule de Stirling et le résultat exact qui serait atteint par le calcul patient des combinaisons et la mise en formule de leur résultat, apparaît énorme et elle est d'autant plus grande que le nombre d'épreuves est grand. Mais, par contre, l'erreur relative sera extrêmement petite, d'autant plus petite que le nombre absolu sera grand; cette erreur relative est la seule qui nous importe, et comme la probabilité serait suffisamment connue à un centième près, la formule de Stirling qui nous assure d'un degré de précision bien supérieur et tendant vers zéro, répond à toutes les exigences possibles.

Pour mettre en nombre la formule de Stirling, il faut remplacer n par le nombre auquel on s'arrête, ϵ et π par leurs valeurs indiquées plus haut. Le calcul se fait par loga-

---

(1) STIRLING a publié sa formule en 1730, mais elle n'a été établie avec rigueur qu'au XIXe siècle. La démonstration en a été faite par Tchebichef, en 1846, et elle a été développée par Rouché en 1890. M. Mansion l'a généralisée en 1908 et 1911; notre savant compatriote, dans la note II annexée au « Cours de calcul des probabilités » de E.-J. Boudin, a montré comment on peut exposer la théorie de Tchebichef d'une manière entièrement élémentaire.

(Cfr. EMMANUEL-JOSEPH BOUDIN : *Leçons de calcul des probabilités* faites à l'Université de Gand de 1846 à 1890, publiées avec des notes et des additions par Paul Mansion, professeur à l'Université de Gand, membre de l'Académie Royale de Belgique. Paris, Gauthier-Villars, 1916, note II (formule de Stirling), p. 244.)

rithmes. La partie entière du logarithme Sn indique le nombre de chiffres, auquel il faut ajouter 1, dont se compose le nombre cherché; en remontant du logarithme au nombre, par les tables, on trouve l'expression approchée, dont l'erreur est d'autant plus petite que le nombre exact est élevé (1).

406. L'application de la formule de Stirling comporte la multiplication par 2 du nombre limite, celui auquel on désire s'arrêter ($\sqrt{2\pi n}$). Dans le calcul de Pn, on arrive à une expression qui, par suite des simplifications entre le numérateur et le dénominateur, se réduit à $\sqrt{\frac{2}{\pi}}$. Cette constante a pour nombre 0.797816 et son logarithme est 9.9019401; le nombre 0.797816 multiplié par le nombre de termes plus 1, compris dans l'expression fractionnaire exprimant la probabilité, nombre lui-même divisé par 100, donne la valeur de la probabilité cherchée. La probabilité de l'écart de devenir de plus en plus faible est corrélative de la probabilité de séries de plus en plus longues; ces probabilités qui diffèrent de moins en moins et se rapprochent toujours davantage de la limite $\sqrt{\frac{2}{\pi}}$ peuvent être calculées aisément à l'aide d'une formule tirée du théorème de Bernouilli, formule qui elle-même est remplacée par une seule table numérique qui, une fois calculée, peut trouver son application dans tous les cas.

Voici cette table telle qu'elle est reproduite par M. Carvallo d'après Houël (Recueil de formules et de tables numériques), à la fin de son ouvrage sur le calcul des probabilités.

(1) On trouvera dans Carvallo, p. 17, un exemple complet de la mise en nombre de la formule de Stirling.

TABLE DES PROBABILITÉS DES ÉCARTS D'APRÈS LA FORMULE RÉDUITE :

$$P = \frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2}\,dt \qquad (1)$$

t	0	1	2	3	4	5	6	7	8	9	d
0.0	0 0000	0113	0226	0338	0451	0564	0676	0789	0901	1013	112
1	1125	1236	1348	1459	1569	1680	1790	1900	2009	2118	109
2	2227	2335	2443	2550	2657	2763	2869	2974	3079	3183	103
3	3286	3389	3491	3593	3694	3794	3893	3992	4090	4187	97
4	4284	4380	4475	4569	4662	4755	4847	4937	5028	5117	88
0.5	0 5205	5292	5379	5465	5549	5633	5716	5798	5879	5959	80
6	6039	6117	6194	6270	6346	6420	6494	6566	6638	6708	70
7	6778	6847	6914	6981	7047	7112	7175	7238	7300	7361	60
8	7421	7480	7538	7595	7651	7707	7761	7814	7867	7918	51
9	7969	8019	8068	8116	8163	8209	8254	8299	8342	8385	42
1.0	0.8427	8468	8508	8548	8586	8624	8661	8698	8733	8768	34
1	8802	8835	8868	8900	8931	8961	8991	9020	9048	9076	27
2	9103	9130	9155	9181	9205	9229	9252	9275	9297	9319	21
3	9340	9361	9381	9400	9419	9438	9456	9473	9490	9507	16
4	9523	9539	9554	9569	9583	9597	9611	9624	9637	9649	12
1.5	0.9661	9673	9684	9695	9706	9716	9726	9736	9745	9755	8
6	9763	9772	9780	9788	9796	9804	9811	9818	9825	9832	6
7	9838	9844	9850	9856	9861	9867	9872	9877	9882	9886	5
8	9891	9895	9899	9903	9907	9911	9915	9918	9922	9925	3
9	9928	9931	9934	9937	9939	9942	9944	9947	9949	9951	2
2.0	0.9953	9970	9981	9989	9993	9996	9998	9999	9999	0000	—

Ainsi, d'après cette table, pour connaître la probabilité que l'écart réduit ne dépasse pas un certain nombre, on commence par chercher le premier chiffre de ce nombre

(1) Les Allemands, dit M. MANSION, *loc. cit.*, note III, p. 251, nomment l'intégrale J *intégrale de Gauss*, mais Gauss lui-même l'appelle toujours *intégrale de Laplace*. L'illustre mathématicien français l'a en effet étudiée dès 1778, dans un *Mémoire sur les probabilités* publié en 1781 par l'Académie des Sciences de Paris.

dans la première colonne, puis on cherche le second chiffre dans une des colonnes numérotées de 0 à 9 et on s'arrête en face de la ligne occupée par le premier chiffre : le nombre écrit dans la seconde colonne à cet endroit est la probabilité que l'on cherche (1).

La table des écarts réduits peut se traduire graphiquement. Il suffit de porter sur l'axe des abcisses l'échelle de l'écart réduit t et sur l'axe des ordonnées les divisions de l'unité de 0 à 1, en donnant une importance égale aux divisions portées sur chacun des deux axes. Ainsi le point 0.1 des ordonnées correspondant au point 0.1 des abcisses a pour valeur 0.1236; on peut indifféremment construire le graphique à l'aide de la table ou remonter du graphique à la table numérique. L'utilisation de la table suppose l'application de la règle de l'écart-étalon; on la formule ainsi : l'écart-étalon est celui qui correspond au nombre 1 de la table de Bernouilli et au nombre 1 du graphique procédant de la table; la probabilité d'avoir un écart inférieur à l'écart-étalon est 0.84. La formule du calcul pour cet écart est

$$e = \sqrt{\frac{2p(1-p)}{n}} \qquad (86)$$

Ensuite, si l'on veut connaître la probabilité que l'écart ne dépasse pas un nombre donné x, on divisera l'écart donné par l'écart-étalon. Le quotient correspond à l'écart réduit t que l'on trouve en tête de la première colonne de la table et ce calcul permet enfin l'utilisation de celle-ci.

407. Jusqu'à présent, nous nous sommes efforcé de condenser en quelques pages les considérations d'ordre mathématique qui sont le commentaire nécessaire du théorème

(1) On trouvera une table plus détaillée dans l'ouvrage de M. de Montessus : *Leçons élémentaires sur le calcul des probabilités*, p. 44 et une table très réduite dans Borel : *Éléments de la théorie des probabilités*, p. 54.

de Bernouilli et les tables graphique et numérique qui en sont la traduction. Il nous reste à montrer comment ce hors d'œuvre a son utilité à l'égard des statistiques.

Le théorème de Bernouilli définit les conditions du hasard; or, la statistique, par ses vérifications *a posteriori*, permet d'établir la loi des phénomènes qu'elle étudie; cette loi est-elle conforme à la loi du hasard? La vérification de ce point est d'un puissant intérêt, elle nous conduit à une classification des faits conforme à une notion essentiellement scientifique. Pour cette vérification, on fera la moyenne des données (mesures, nombres) de la statistique et l'on calculera les écarts de chaque donnée à cette moyenne; si les écarts classés par grandeur se groupent d'une façon sensiblement analogue à la courbe dérivée de la formule de Bernouilli, l'assimilation du phénomène avec les données résultant du hasard est permise.

Les mathématiciens se sont montrés parfois très durs à l'égard de la statistique et des statisticiens. Habitués à raisonner sur des formules rigides, ils ne peuvent tolérer ce qu'il peut y avoir d'imprécis dans certains résultats statistiques qui sont plus des approximations qu'une véritable détermination de la vérité. Les sciences de pur raisonnement ont cet avantage sur les sciences d'observation qu'elles ne dépendent presque pas des contingences; au contraire toute observation se heurte à de multiples difficultés : erreurs des instruments; maladresse de l'observateur; nécessité de recourir, comme dans les statistiques, à de très nombreux collaborateurs dont le zèle, l'intelligence et la probité scientifique ne peuvent être garantis d'une manière absolue; manque de sincérité des personnes interrogées; moyens financiers et matériels insuffisants...

Par le fait même, les sciences de raisonnement peuvent présenter des résultats plus précis, plus exacts que les autres disciplines qui s'appuient sur des observations; non contents de cet avantage, les mathématiciens ou certains d'entre eux ont voulu identifier la valeur d'une statistique

quelconque avec sa conformité à la loi du hasard. Nous pensons que c'est une erreur et qu'il convient de se montrer plus large et plus précis à la fois; certains résultats statistiques sont gouvernés par la loi du hasard, c'est incontestable, et à ceux-là le théorème de Bernouilli est entièrement applicable, mais à côté de ces cas spéciaux, il est une multitude de résultats provenant d'une observation scientifique de valeur, qui ne sont pas soumis à la loi dont il s'agit. La statistique peut être bien faite et n'avoir rien de commun avec la répartition des chances résumée dans la table de Bernouilli. La démonstration la plus évidente de cette vérité se trouve dans l'exemple de statistiques judiciaires: peut-on ou ne peut-on pas déterminer la probabilité qu'un innocent a d'être condamné en justice? Cette question, encore qu'elle paraisse étrange à première vue, a intéressé un grand nombre de mathématiciens, et non des moindres : Condorcet, Laplace, Poisson, Cournot... Chacun d'eux a proposé sa solution : sans descendre à l'analyse mathématique de leurs formules, on peut nier l'exactitude de celles-ci parce que la probabilité d'un jugement équitable n'est pas et ne peut pas être connue. Ce n'est que par des sophismes de raisonnement et de calcul qu'on est parvenu à donner une valeur à cette probabilité (1); toujours il a fallu admettre une proposition non démontrée; Condorcet assi-

---

(1) M. Mansion, après avoir cité les paroles de Bertrand « on peut peser le cuivre et le donner pour or, la balance est sans reproche », paroles qui répondaient au mot célèbre de Stuart Mill (l'application du calcul des probabilités aux décisions judiciaires est le scandale des mathématiques) ajoute que cet argument de Bertrand excuse les mathématiques mais n'excuse pas les mathématiciens d'avoir tenté de soumettre au calcul des probabilités des évènements qu'il est évidemment impossible de ranger sous une loi, si peu strictement approximative qu'on la suppose. Au moins avons-nous cette consolation, continue M. Mansion, que c'est un mathématicien aussi, Bertrand lui-même qui a définitivement banni des traités ce chapitre humiliant sur la probabilité des jugements.

Cfr. Mansion, *loc. cit.* Note XIII, p. 314.

milait les jugements équitables à des boules blanches tirées d'une urne contenant, à côté de boules noires, une certaine proportion de boules blanches. Mais quelle proportion?

Si le nombre de boules blanches est plus grand que celui des boules noires, il doit y avoir pour l'innocent plus de chances de voir résoudre son cas par un jugement favorable que d'être condamné, et cette chance augmente à mesure que le nombre de tirages est plus considérable : aussi un innocent aurait-il, dans cette hypothèse, tout avantage à faire prononcer sur son procès tous les juges du pays. Le cas inverse se présenterait si la composition de l'urne était dans le sens contraire.

On a aussi imaginé d'obtenir la valeur de la probabilité en se basant sur le nombre de jugements réformés. Ce système pèche contre la logique, car rien ne nous dit que les juges d'appel auront une appréciation plus juste que celle des magistrats de première instance; puis il y a des décisions judiciaires réformées pour simple vice de forme, et celles dont l'accusé, pour un motif quelconque, et qui n'est pas toujours un aveu, n'interjette pas appel; bref, il faut renoncer à rechercher une probabilité qui ne peut pas être définie (1).

L'exemple ci-dessus prouve qu'il ne suffit pas qu'une statistique soit aussi bien faite qu'on le désire pour que le théorème de Bernouilli lui soit applicable.

408. Après ces considérations d'ordre général, approchons de plus près le terrain d'application du théorème de Bernouilli aux faits statistiques et demandons-nous quels

---

(1) M. Mansion, *loc. cit.* cite ces paroles de Bertrand : « L'indépendance des tirages (dans le cas du tirage de boules d'une urne qui en contient des blanches et des noires) est supposée; les urnes, dans les calculs, échappent à toute influence commune. Les juges, au contraire, s'éclairent les uns les autres; les mêmes faits les instruisent, les mêmes témoignages les troublent, la même éloquence les égare, c'est sur les mêmes considérants qu'ils font reposer la vérité ou l'erreur. L'assimilation est impossible. » (Bertrand, p. XIV.)

sont les problèmes auxquels ce théorème réserve une solution. Nous avons vu plus haut que parmi les desiderata principaux de la science, on pouvait faire figurer la comparaison des résultats de l'expérience avec les données de la théorie. Ainsi, il est important en statistique de comparer la distribution des écarts annuels d'un phénomène à la distribution théorique des écarts accidentels. Ce problème avait été proposé en France à un concours pour le recrutement de statisticiens officiels en appliquant les données à la proportion observée de naissances masculines, sachant que la moyenne de 26 années couvertes par l'observation est pour les

Nés vivants	378,241 garçons	356,208 filles	734,449 total
Mort-nés	10,319 garçons	7,637 filles	17,956 total

Cet exemple a été choisi par M. Carvallo pour exposer la méthode à suivre dans l'étude de la comparaison entre les écarts observés et les écarts théoriques. Nous suivrons ici les explications de M. Carvallo, en les abrégeant le plus souvent et en les développant dans d'autres cas (1).

Nous commençons par admettre que tout le matériel statistique est réuni et que pour chacune des 26 années sur lesquelles portent les observations on connaisse le nombre d'enfants nés vivants ou mort-nés, garçons et filles. Le calcul de la moyenne générale se fait d'après la règle de la moyenne arithmétique simple, méthode directe, dont nous rappelons la formule :

$$M = \frac{1}{N} \Sigma (X) \tag{10}$$

Cette moyenne générale étant connue, il y a lieu ensuite de calculer les écarts de chaque donnée annuelle à la moyenne d'ensemble. On étudie ces écarts en calculant tout d'abord la fréquence de l'apparition du phénomène pour

(1) CARVALLO : *Le calcul des probabilités et ses applications*. Paris, Gauthier-Villars, 1912, pp. 41-62.

chaque année et pour la moyenne formée de l'ensemble. L'unité marquant la certitude, la probabilité *a posteriori p* s'exprimera par une fraction décimale; celle-ci sera comparée à la moyenne et l'écart, selon qu'il sera positif ou négatif, sera précédé du signe + ou du signe —. L'écart avec la moyenne est parfois très faible, parfois nul.

Dans l'exemple donné l'écart le plus considérable est de 18 unités de l'ordre du dix-millième; c'est un écart négatif qui s'est produit la 22e année; par contre, il y a trois années où il y a coïncidence absolue entre la probabilité théorique et la probabilité *a posteriori.*

Le classement des écarts par ordre de grandeur est ensuite à effectuer. Le chiffre de 18 unités de l'ordre du dix-millième comme écart maximum suggère immédiatement le chiffre de deux unités du même ordre pour classer les écarts d'après leur grandeur; on pourrait aussi prendre le chiffre de trois unités, mais ce choix aurait pour conséquence de réduire le nombre de classes à 6 au lieu de 9, ce qui serait de nature à rendre plus confuse l'étude complète du phénomène. Il est donc formé un tableau comprenant 9 divisions de 2 en 2 dix-millièmes; en regard de la première colonne réservée à ces divisions (2, 4, 6, 8, ... 18) l'étudiant inscrira le nombre d'écarts qui restent dans la limite indiquée, c'est-à-dire qui ne sont pas supérieurs à 2 dix-millièmes, etc.; on fera attention à deux choses : c'est que les données indiquant un écart zéro doivent entrer en ligne de compte et qu'à la deuxième division (4) on compte non seulement les écarts compris entre 2,1 et 4, mais aussi ceux faisant partie de la première division 0 à 2. Le nombre de ces écarts est appelé *nombre de distribution* correspondant à l'écart X et est désigné par la lettre Φ (X). Les nombres ainsi obtenus sont inscrits en regard de chaque division et forment la seconde colonne du tableau. Une troisième colonne est obtenue en y portant la différence entre le 1er et le 2e chiffre de la seconde colonne, entre le 2e et le 3e et ainsi de suite;

cette colonne, dans le cas pris pour exemple, compte 8 divisions, c'est-à-dire une de moins, comme il est naturel, que les deux colonnes précédentes.

409. Cette phase des opérations peut être considérée comme la phase préparatoire ; elle met le matériel en état d'être utilisé dans le but qu'on se propose. Il faut maintenant étudier la question de savoir quelle serait la distribution théorique des écarts, si l'on admet que l'apparition du phénomène est réglée comme si elle était assimilable à la sortie d'une boule blanche ou noire d'une urne dont la composition resterait constante, la probabilité de la naissance d'un garçon étant celle indiquée comme probabilité *a posteriori;* à savoir : 0.5150. Déterminons tout d'abord la position qui correspond à la longueur 1 de l'échelle de la courbe ou au nombre 1 de la table de Bernouilli, sachant que la probabilité de l'événement est 0.5150 et que le nombre des épreuves, autrement dit le nombre total de naissances, est 734,449. La valeur que nous cherchons est celle de l'écart-étalon dont la formule est :

$$e = \sqrt{\frac{2p(1-p)}{n}} \qquad (86)$$

or,

$$1 - p = 1 - 0.5150 = 0.4850$$

$$2p(1-p) = 2 \times 0.5150 \times 0.4850 = 0.49955.$$

$$\frac{2p(1-p)}{n} = \frac{0.49955}{734449} = 0.00000068017$$

$$e = \sqrt{\frac{2p(1-p)}{n}} = 0.000824 \quad (1)$$

Calculons ensuite la valeur réduite t, à l'aide de laquelle on peut utiliser la table de Bernouilli. En regard de t se

(1) Et non 0.000816. valeur calculée par M. CARVALLO, *op. cit.*, p. 44. Cette correction modifie les résultats subséquents obtenus par M. Carvallo, notamment le tableau de distribution, p. 45 du même ouvrage.

trouve dans la table le nombre P (t) exprimant la fréquence théorique à comparer à la fréquence observée. Or,

$$t = \frac{x}{e} = \frac{0.0002}{0.000824} = 0.2427 \quad (1)$$

pour le premier écart, c'est-à-dire le quotient de la division de l'écart brut x par l'écart-étalon.

Par écart brut, nous entendons le chiffre qui exprime les grandeurs des écarts rangés par classe. Pour le second cas de la série

$$t = \frac{0.0004}{0.000824} = 0.4854$$

Les nombres de distribution $\Phi\ (x)$ doivent être divisés par le nombre des années observées (N). Ce résultat a pour symbole $\varphi\ (t)$ fréquences qui correspondent aux valeurs échelonnées de t.

410. L'ensemble de ces données compose le tableau de distribution, à propos duquel il y a lieu de rappeler les valeurs et formules suivantes :

$n = 734449$ $\qquad$ $e = 0.000824$

$\mathrm{N} = 26$ $\qquad$ $t = \frac{x}{e}$ (premier cas $= \frac{0.0002}{0.000824} = 0.2427$

$p = 0.5150$ $\qquad$ $\varphi t = \frac{\Phi\ (x)}{\mathrm{N}} \left(\text{premier cas} = \frac{8}{26} = 0{,}3077 \text{ ou } 0{,}308\right)$

(1) Et non 0,245, correction qui est la conséquence de la remarque faite à la note de la page précédente.

Tableau des écarts.

ANNÉES	Fréquence des naissances masculines rapportées à l'unité	Écarts positifs	Écarts négatifs	Écarts nuls
1	0.5164	0.0014		
2	0.5166	0.0016		
3	0.5150			0
4	0.5157	0.0007		
5	0.5151	0.0001		
6	0.5156	0.0006		
7	0.5159	0.0009		
8	0.5155	00.005		
9	0.5161	0.0011		
10	0.5140		0.0010	
11	0.5150			0
12	0.5149		0.0001	
13	0.5140		0.0010	
14	0.5145		0.0005	
15	0 5148		0.0002	
16	0.5147		0.0003	
17	0.5152	0 0002		
18	0.5145		0.0005	
19	0.5138		0.0012	
20	0.5144		0.0006	
21	0.5157	0.0007		
22	0.5132		0.0018	
23	0.5146		0.0004	
24	0.5151	0.0001		
25	0.5150			0
26	0.5147		0.0003	

M 0.5150 (p).

Classement des Écarts

x	$\Phi(x)$	Différences
2	8	
		3
4	11	
		5
6	16	
		2
8	18	
		3
10	21	
		2
12	23	
		1
14	24	
		1
16	25	
		1
18	26	

Distribution et Réduction.

x	$\Phi(x)$	t	$\varphi(t)$
0.0002	8	0.2427	0.308
0.0004	11	0.4854	0.423
0.0006	16	0.7281	0.615
0.0008	18	0.9708	0.692
0.0010	21	1.2135	0.808
0.0012	23	1.4563	0.885
0.0014	24	1.6990	0.923
0.0016	25	1.9417	0.962
0.0018	26	2.1844	1.000

411. Les premiers calculs étant ainsi effectués, il y a lieu de recourir à la comparaison graphique entre la courbe théorique et la courbe représentative des écarts observés.

Commençons par construire la courbe théorique. Dans ce but, nous utilisons la table de Bernouilli dont nous avons reproduit plus haut un modèle réduit. (Cfr. n° 396.) Le lecteur pourra utiliser s'il le désire, une table plus détaillée s'il veut arriver à une approximation plus exacte (1). Sur l'axe horizontal des abcisses nous portons, à intervalles égaux, les divisions correspondantes aux nombres *t* de la table, à savoir : 0,0; 0,1; 0,2; 0,3; etc., jusqu'au nombre 4,80 dont la valeur correspondante est 0,99999999999, mais pratiquement on peut s'arrêter au nombre 2 dont la valeur est déjà 0,9953223.

Sur l'axe vertical des ordonnées, nous portons des divisions égales dont l'ensemble constitue l'échelle des probabilités, allant de 0,0 à 1, ce dernier chiffre marquant la certitude.

En lisant la table, nous voyons que le point 0.1 = 0.1125, le point 2 a pour valeur correspondante 0.2227, le point 0.9 sera représenté par la valeur 0.7969, etc. En réunissant tous ces points par un trait, nous obtenons la courbe théorique de probabilité; elle est représentée sur le graphique ci-après par le trait plein. Nous avons maintenant à reporter sur le même graphique les fréquences observées; consultons la table de réduction (Cfr. n° 410) : nous y trouvons les valeurs de *t* et de φ (*t*). Les valeurs de *t* seront portées sur l'échelle des abcisses et les valeurs de φ (*t*) mesurées sur l'axe des ordonnées par une verticale menée de chaque point *t*; donc au point 0.2427 situé sur l'axe horizontal correspond une valeur 0.308 immédiatement au-dessus et ainsi de suite. Les points déterminés de la sorte sont inscrits sur le graphique au moyen d'un petit cercle. Pour relier ces points entre eux, on trace une courbe interpolatrice, car la position des points n'est pas

(1) On en trouvera une très détaillée dans l'ouvrage cité de M DE MONTESSUS, p. 44-47.

telle qu'ils puissent être tous unis par une courbe. Le graphique est alors construit et l'on dispose des éléments nécessaires à la comparaison de la courbe théorique et de celle des fréquences observées.

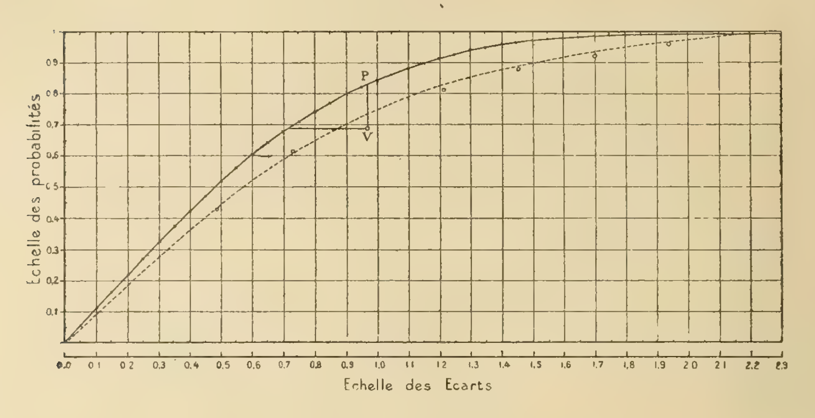


Fig. 43.

412. A part un cas isolé, les points de fréquence observée sont restés en dessous de la courbe théorique. Il n'y a donc pas assimilation complète entre la loi du hasard et la loi de nature physiologique, qui règle la fréquence des naissances masculines parmi les individus nés vivants. Mais quelle est la mesure de cet écart? On peut la concevoir comme exprimant la chance d'atteindre, au moyen d'un tirage au sort, le point correspondant sur la courbe théorique. M. Carvallo partant de cette idée a établi une méthode simple d'exprimer l'écart en question.

Prenons sur la courbe des fréquences observées le point le plus proche de l'unité de l'échelle des écarts, et appelons-le V. La valeur de ce point est $t = 0.9708$, $\varphi(t) = 0.692$, d'après le tableau de distribution inséré au n° 410.

En tirant de ce point une ligne horizontale vers la gauche, cette ligne coupe la courbe théorique au point 0.69 qui correspond à la mesure $t = 0.72$. C'est donc à l'écart théorique $t = 0.72$ et non à

l'écart observé $t = 0.97$ que devrait correspondre la fréquence $\varphi(t) = 0.692$. Donc les écarts observés sont trop grands. Mais dans quelle mesure le sont-ils?

Prenons d'abord la différence entre la fréquence du point V sur la courbe des observations et la probabilité correspondante P que nous obtenons en menant une verticale de V sur la courbe des observations, à P sur la courbe théorique; la valeur cherchée est inscrite en regard sur l'échelle des probabilités.

$$V = 0.9708 \qquad\qquad P = 0.8299$$

Nous obtenons ces valeurs indifféremment par le graphique ou par la table.

D'après la position de V sur la courbe observée, nous avons la valeur

$$t = 0.97$$

et la probabilité de sortir un écart inférieur est donnée par la table qui indique

$$P(t) = 0.8299$$

Donc,

$$P - \varphi = 0.83 - 0.69 = 0.14$$

Les années observées sont au nombre de 26 (N). L'écart étalon sera

$$E = \sqrt{\frac{2P(1-P)}{26}} = \sqrt{0.0108} = 0.10392$$

L'écart réduit

$$T = \frac{X}{E} = \frac{0.14}{0.10392} = 1.34.$$

Or, d'après la table, à $t = 1.34$ correspond la valeur 0.9419.

Il y aurait donc à parier 94 contre 6 que le point V serait au-dessus de la position qu'il a sur le graphique s'il y avait assimilation complète avec la loi du hasard ; en conséquence 0.94 marque la défaveur jetée sur la série par la position du point V.

III. — La loi des erreurs.

413. La loi des erreurs, qu'on désigne parfois sous le nom de loi de Gauss, prend une expression mathématique assez compliquée que nous définirons plus loin, mais les données de l'expérience suffisent pour en comprendre la nature et la portée.

C'est un fait universellement connu et admis, qu'aucune mesure obtenue par l'observation n'est rigoureusement exacte ; tantôt ce sont les moyens dont nous disposons qui sont insuffisants pour atteindre l'exactitude absolue, tantôt ce sont les conditions du milieu où se fait l'observation qui sont défavorables, tantôt ce sont les qualités de l'observateur lui-même qui ne sont pas à la hauteur de sa tâche. Imperfection des instruments d'investigation, erreurs de l'observateur, conditions défectueuses du travail se combinent ainsi pour éloigner de la vérité absolue, dans une mesure plus ou moins notable, le résultat de notre travail. Un exemple ne sera pas inopportun pour préciser ce qui vient d'être dit. Nous l'emprunterons à Quetelet, qui fut un précurseur dans cette question en étendant la loi des erreurs de son domaine mathématique à l'observation des plantes et des êtres humains. Dans son célèbre ouvrage sur la théorie des probabilités, l'illustre statisticien et sociologue s'exprime comme suit au sujet des erreurs qu'on relève dans les observations astronomiques (1) : « Un ob-

---

(1) QUETELET, *Lettres sur la théorie des probabilités*, lettre XIX. Bruxelles, Hayez, 1846, pp. 124-125.

servateur peut assigner la position d'un astre sans avoir à craindre des erreurs qui s'élèvent à plus de trois à quatre secondes en arc, c'est-à-dire que la distance dont il peut se tromper équivaut, au plus, à la largeur de la petite bande du ciel que nous cacherait un fil, tendu à plusieurs pieds de distance devant nos yeux. Cependant un grand nombre de causes peuvent donner naissance à cette erreur, et nous nous plaçons ici dans l'hypothèse la plus défavorable; nous supposons qu'elles tendent toutes, dans le même sens, à donner une valeur soit trop grande, soit trop petite. Ainsi, quelque précis que soit l'instrument, il n'est pas parfait dans toutes ses parties : quelles que soient l'adresse et l'expérience de l'observateur, son coup d'œil n'est pas infaillible; l'air peut être dans des circonstances plus ou moins défavorables : nous ne voyons les astres que du fond de l'atmosphère où nous sommes plongés; et, à cause des réfractions, ils ne sont réellement pas dans les lieux où nous les apercevons. »

On pourrait multiplier ces exemples. Il n'est pas inutile de faire remarquer que les erreurs accidentelles d'observation peuvent chacune provenir de plusieurs causes et non d'une cause unique. Ainsi, écrit M. Th.-W. Wright, en lisant un angle au théodolite, l'erreur trouvée peut être le résultat de la construction imparfaite de l'instrument ou du manque de précision dans la façon dont l'observateur manie son instrument. Chacune de ces causes peut être le résultat d'influences diverses : ainsi la première nommée englobe les erreurs de collimation de niveau, etc. (1).

Pour chacun de ces cas, nous supposons que l'observateur est de bonne foi et que les instruments dont il se sert ne sont pas construits de façon à donner une erreur systématique. Il faut que les erreurs soient accidentelles; c'est la condition essentielle pour qu'on se trouve dans le do-

---

(1) Wright (Thomas-Walton). *The adjustment of observations* New York, 1906, p. 17.

maine du calcul des probabilités; aussitôt qu'on s'en éloigne, dès qu'on suppose que des erreurs systématiques interviennent, on est en dehors du calcul et de ses règles.

Les erreurs accidentelles ne sont-elles soumises à aucune loi? Se produisent-elles indifféremment dans tous les cas? Il suffit d'un instant de réflexion pour s'apercevoir que rien ne semble moins probable. Il est aisé de se tromper d'une petite quantité dans une observation : le plus habile observateur peut mal lire un résultat, laisser échapper l'un ou l'autre élément d'une importance relative. Par contre, les erreurs grossières sont rares; il faut imaginer, pour les rendre explicables, une série de circonstances bizarres qui ne peuvent se trouver réunies que rarement. Un bon nombre de petites erreurs, quelques rares erreurs sérieuses, tel paraît bien être le bilan d'une série d'observations faites avec soin. D'autre part il y aura toujours des erreurs; il n'est pas possible de les éliminer toutes : « L'expérience prouve, dit M. Gruey, que les diverses mesures d'une même grandeur, faites soigneusement avec toute la précision dont la méthode employée est susceptible et purgées de toutes les erreurs systématiques, présentent encore de petites différences. »

Enfin, si nous connaissons la valeur centrale de toutes les erreurs, celle qui est également éloignée de la plus petite et de la plus grande, ces mots étant entendus dans le sens de l'arithmétique, nous devons reconnaître qu'il y a une probabilité égale à ce que les erreurs se produisent dans l'un ou l'autre sens et prennent, par rapport à cette valeur centrale, le signe positif ou le signe négatif. Ce sont précisément ces caractères qui sont mis en évidence par la formule de Gauss. Quant à la démonstration, ce sont, comme le dit M. Borel, les raisons les plus simples qui sont les meilleures (1).

---

(1) Emile Borel, *Eléments de la théorie des probabilités*. Paris, Hermann et Fils, 1909, p. 129.

C'est à tel point que ce mathématicien de talent estime inutiles, en présence de la valeur démonstrative des vues logiques, les développements mathématiques donnés à la question ; il ne lui semble pas que les résultats obtenus dans cette voie aient une importance en rapport avec l'effort analytique qu'ils exigent. L'hypothèse de l'ingénieur allemand G. Hagen (1797-1884), dans un ouvrage élémentaire qui contient un exposé original de la théorie des erreurs, est aujourd'hui admise de plus en plus par les auteurs qui écrivent sur le calcul des probabilités. On peut admettre, dit Hagen, que dans chaque espèce d'observations, il existe un certain nombre indéterminé d'erreurs élémentaires, indépendantes les unes des autres, qui, prises d'une manière absolue, sont de même grandeur et peuvent être indifféremment positives ou négatives. C'est la somme algébrique de ces erreurs élémentaires qui forme, dans chaque cas particulier, la véritable erreur existante (1)

414. Avant tout, nous devons revenir en deux mots sur le sujet de la moyenne. Nous avons montré que de toutes les valeurs différentes que prennent les observations faites d'une quantité unique, la moyenne était la plus exacte et que quand le nombre des valeurs observées est très grand, la moyenne arithmétique était la valeur véritable. D'autre part, après avoir précisé la notion de l'écart, nous avons montré : *a*) que la somme algébrique des écarts est égale à zéro ; *b*) que la somme des carrés des écarts d'après la moyenne arithmétique constitue un minimum. L'étude de la moyenne nous a aussi conduit à cette conclusion que si l'on dispose d'un matériel dont les éléments ont une valeur identique sous le rapport de l'observation, on doit trouver que les erreurs en trop ou par défaut se partagent d'une manière égale autour de la moyenne, c'est-à-dire

(1) Cfr. Boudin-Mansion, *Cours*, *loc. cit.*, pp. 154 et 289.

qu'elles présentent une égale probabilité. Nous avons dit que l'expérience démontre, comme il est naturel, la rareté des erreurs considérables et la fréquence des petites erreurs. Les unes comme les autres se rangent autour de la moyenne; leur nombre augmente, de part et d'autre, à mesure qu'on se rapproche de zéro, de sorte qu'il est permis de dire que l'arrivée des erreurs en nombre plus ou moins grand est une fonction de l'erreur elle-même. M. Th.-W. Wright (1) a fait l'exposé suivant de la genèse de la formule de la loi des erreurs; nous le reproduisons intégralement à cause de sa clarté :

Appelons $f\ (\Delta)$ la probabilité qu'une erreur prenne place entre 0 et Δ ; et désignons par q la probabilité qu'une autre erreur se rencontre entre Δ et $\Delta + d\,\Delta$. Nous avons :

$$q = f(\Delta + d\,\Delta) - f(\Delta) = f'\,(\Delta)\,d\,\Delta = \varphi\,(\Delta)\,d\,\Delta. \qquad (87)$$

La fonction $\varphi\ (\Delta)$ est appelée loi de distribution des erreurs.

Si nous voulons exprimer la probabilité qu'une certaine erreur tombe entre les limites a et b de la courbe, nous avons à calculer la probabilité $\varphi\ (\Delta)\ d\ \Delta$ étendue aux limites sus indiquées, ce qui nous donne l'intégrale :

$$\int_a^b \varphi\,(\Delta)\,d\,\Delta. \qquad 88)$$

La probabilité qu'une erreur donnée ne dépasse pas la limite a s'exprime ainsi :

$$\int_{-a}^{+a} \varphi\,(\Delta)\,d\,\Delta. \qquad (89)$$

La probabilité totale de l'arrivée simultanée de toutes les erreurs possibles est la somme des probabilités calcu-

(1) Th. W. Wright, *loc. cit.*, pp. 13 et suivantes.

lées pour chaque erreur séparément. Désignons-la par Q. La probabilité de la rencontre de l'erreur Δ_1 est $\varphi(\Delta_1)$ $d\,\Delta_1$ et ainsi de suite pour $\Delta_2, \ldots . \Delta_n$; nous avons en conséquence :

$$Q = \varphi(\Delta_1)\,\varphi(\Delta_2)\ldots\ldots\varphi(\Delta_n)\,d\,\Delta_1,\,d\,\Delta_2,\ldots.\,d\,\Delta_n. \qquad (90)$$

Q est un maximum et son logarithme partage la même propriété. En établissant les différences par rapport à x, inconnue dont il faut déterminer la valeur la plus probable, on a :

$$0 = \frac{d\,(\log.\ Q)}{dx} = \frac{\varphi'(\Delta_1)}{\varphi(\Delta_1)}\frac{d\,\Delta_1}{dx} + \frac{\varphi'(\Delta_2)}{\varphi(\Delta_2)}\frac{d\,\Delta_2}{dx} + \ldots . \frac{\varphi'(\Delta_n)}{\varphi(\Delta_n)}\frac{d\,\Delta_n}{dx}$$

$$= \frac{\varphi'(\Delta_1)}{\Delta_1\,\varphi(\Delta_1)}\,\Delta_1 + \frac{\varphi'(\Delta_2)}{\Delta_2\,\varphi(\Delta_2)}\,\Delta_2 + \ldots\ldots \frac{\varphi'(\Delta_n)}{\Delta_n\,\varphi(\Delta_n)}\,\Delta_n \qquad (91)$$

or,

$$\frac{d\,\Delta_n}{dx} = 1.$$

Nous savons que si le nombre des erreurs d'observation est très grand, ces erreurs se compensent et ont pour expression finale zéro.

Nous avons donc :

$$\frac{\varphi'(\Delta_1)}{\Delta_1\,\varphi(\Delta_1)} = \frac{\varphi'(\Delta_2)}{\Delta_2\,\varphi(\Delta_2)} = \ldots \frac{\varphi'(\Delta_n)}{\Delta_n\,\varphi(\Delta_n)} = K.\ \ldots \qquad (92)$$

A toute valeur arbitraire Δ, correspond donc l'expression générale :

$$\frac{\varphi'(\Delta)}{\Delta\,\varphi(\Delta)} = K.$$

Par intégration et simplification, on a :

$$\varphi(\Delta) = c\,e^{\frac{1}{2}K\,\Delta^2} \qquad (93)$$

expression dans laquelle e est la base du système népérien de logarithmes et c est une constante à déterminer.

Puisque Q est un maximum, $\frac{d^2\,Q}{dx^2}$ ou $\frac{d^2\,(\log.\,Q)}{dx^2}$ doit être une expression négative. D'où

$$Q = c^n\, e^{\frac{k}{2}(\Delta_1^2 + \Delta_2^2 + \dots)}\, d\,\Delta_1\, d\,\Delta_2 \dots$$

$$d\,\frac{(\log.\,Q)}{dx} = K\,(\Delta_1 + \Delta_2 + \dots)$$

$$d^2\,\frac{(\log.\,Q)}{dx^2} = K\,(\Delta_1 + \Delta_2 + \dots)$$

En conséquence, puisque n est positif, K doit être négatif et prenant la valeur $\frac{1}{2}$ K = —h², nous avons

$$\varphi\,(\Delta) = c\, e^{-h^2\Delta^2} \qquad (94)$$

ce qui donne la loi d'erreur cherchée.

Il faut maintenant déterminer c et h. Il est certain que toutes les erreurs se trouvent entre les limites $+ a$ et $- a$. Par conséquent :

$$c \int_{-a}^{+a} e^{-h^2\Delta^2}\, d\Delta = 1 \qquad (95)$$

Mais comme les valeurs de a sont différentes pour chaque observation particulière, il est préférable de remplacer $- a$ et $+ a$ par le signe général $+ \infty$ $- \infty$ comme désignant les limites extrêmes de l'erreur et d'écrire l'expression générale :

$$c \int_{-\infty}^{+\infty} e^{-h^2\Delta^2}\, d\Delta = 1$$

Il vient que

$$c = \frac{h}{\sqrt{\pi}}. \qquad (96)$$

et la loi des erreurs peut être écrite :

$$\varphi\,(\Delta) = \frac{h}{\sqrt{\pi}}\, e^{-h^2\Delta^2} \qquad (97)$$

h est une mesure de la qualité des observations; Gauss lui a donné le nom de mesure de la précision des observations.

Cette équation a pour expression graphique une courbe symétrique dans laquelle la moyenne, la médiane et le mode coïncident ; elle est la courbe idéale de distribution. Elle porte le nom de courbe des erreurs, ou loi des erreurs.

415. Pour représenter graphiquement la courbe normale des erreurs, on se sert pratiquement de tables calculées à l'avance donnant les ordonnées de la courbe, tables que les mathématiciens connaissent sous le nom de tables de probabilité intégrale. Si l'on considère la courbe comme divisée en 100 parties égales dont les 49 premières divisions. en dessous de 50°, auraient une valeur négative et les 49 dernières divisions, au-dessus de 50°, prendraient une valeur positive, les déviations par rapport à la moyenne s'expriment en valeurs normales, comme suit (1) :

5° = 2.44		60° = 0.38
10° = 1.90		70° = 0.78
20° = 1.25	50° = 0	80° = 1.25
30° = 0.78		90° = 1.90
40° = 0.38		95° = 2.44

La courbe des erreurs présente des particularités dignes de remarque :

1° Elle est symétrique par rapport à l'axe des *y*, c'est-à-dire que son développement de part et d'autre du centre est identique, mais les valeurs qu'elle prend de part et d'autre sont opposées étant de signe contraire ;

2° Elle diminue rapidement, s'abaissant vers l'axe des abcisses, indiquant ainsi que la probabilité de commettre une erreur diminue à mesure que l'erreur devient plus grande ;

3° Elle est asymptotique par rapport à l'axe des abcisses, c'est-à-dire qu'elle s'en rapproche constamment sans se

(1) Francis GALTON, *Natural Inheritance*, table 3, p. 201.

confondre avec lui, ce qui marque la possibilité indéfinie de commettre une erreur.

La table de fréquence des erreurs reproduite page 669 donne, pour 10,000 observations, le nombre théorique des erreurs comprises entre zéro et un nombre quelconque pris comme unité d'écart ou erreur probable.

La première colonne intitulée *m* contient une série de nombres applicables à la grandeur de l'écart probable : ainsi 0.09 signifie une erreur qui ne sera pas supérieure à neuf centièmes de l'écart probable. Il suffit alors de consulter la seconde colonne indiquant le nombre d'erreurs sur 10,000 pour constater qu'à ce point 0.09 correspond la valeur 484 ; sur 10,000 chances, il y a donc 484 chances pour que l'erreur commise ne dépasse pas 9/100 de l'écart probable et de ce nombre 242 seront positives et 242 négatives. Entre deux divisions de la table, toute division intermédiaire a pour valeur la moitié de la différence entre les chiffres indiqués à chacune de ces divisions : ainsi 1.34 (*m*) a pour valeur 6339 ; 1.36 (*m*) correspond au chiffre 6410 ; donc 1.35 prend la valeur $\frac{6410 - 6339}{2} = \frac{71}{2} + 6339 = 6374.5$. Cette règle d'interpolation est applicable à tous les écarts qu'on trouvera dans la table.

416. En utilisant des tables de réduction spéciales, on parvient facilement à comparer la distribution théorique avec la distribution observée.

Nous avons déjà reproduit, d'après M. Carvallo, une de ces tables. M. Benini (1) en a calculé une autre dans laquelle la valeur de $\theta(t)$ est exprimée en pour cent,

(1) Benini, *Principi di statistica metodologica*. Torino, 1906, p. 230. Cette table n'est autre que la table modifiée et simplifiée de l'intégrale :

$$\theta(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-t^2} dt$$

table que l'on trouve sous la forme précédente dans tous les traités de probabilités.

ce qui est d'une grande utilité pour les travaux de statistique pratique. Nous la reproduisons ci-après :

θ (t)	t	θ (t)	t	θ (t)	t	θ (t)	t
1 °/₀	0.0089	26 °/₀	0.2346	51 °/₀	0.4881	76 °/₀	0.8308
2	0.0177	27	0.2440	52	0.4994	77	0.8488
3	0.0266	28	0.2534	53	0.5109	78	0.8673
4	0.0335	29	0.2629	54	0.5224	79	0.8864
5	0.0443	30	0.2725	55	0.5341	80	0.9062
6	0.0532	31	0.2820	56	0.5460	81	0.9267
7	0.0621	32	0.2917	57	0.5580	82	0.9481
8	0.0710	33	0.3013	58	0.5702	83	0.9703
9	0.799	34	0.3111	59	0.5826	84	0.9935
10	0.0889	35	0.3209	60	0.5951	85	1.0179
11	0.0978	36	0.3307	61	0.6078	86	1.0436
12	0.1068	37	0.3406	62	0.6208	87	1.0707
13	0.1157	38	0.3506	63	0.6339	88	1.0994
14	0.1247	39	0.3607	64	0.6473	89	1.1301
15	0.1337	40	0.3708	65	0.6609	90	1.1631
16	0.1428	41	0.3810	66	0.6747	91	1.1988
17	0.1518	42	0.3913	67	0.6888	92	1.2379
18	0.1609	43	0.4017	68	0.7032	93	1.2812
19	0.1700	44	0.4121	69	0.7179	94	1.3299
20	0.1791	45	0.4227	70	0.7329	95	1.3859
21	0.1883	46	0.4333	71	0.7482	96	1.4522
22	0.1975	47	0.4441	72	0.7639	97	1.5345
23	0.2067	48	0.4549	73	0.7800	98	1.6450
24	0.2160	49	0.4659	74	0.7965	99	1.8214
25	0.2253	50	0.47693	75	0.8134		

En multipliant les valeurs de *m* portées à la table de fréquence des erreurs par la constante 0.47693, on obtient les

valeurs de t mentionnées à la table ci-dessus. Pour confronter la distribution observée à la distribution théorique, il faut d'abord calculer l'écart probable et faire cet écart égal à l'unité. Il faut ensuite réduire à la même base la mesure des erreurs ordonnées et chercher son équivalent dans la colonne m de la table de fréquence des erreurs.

Si l'interpolation doit être faite, on la calcule et on lit le nombre dans la colonne en regard de m, en réduisant, si on le désire, à la proportion sur 100.

Les erreurs effectives étant d'abord réduites à leur proportion sur 100, on inscrit en regard les nombres calculés d'après la table et on compare les deux séries. (Voir ci-contre la table de fréquence des erreurs.)

417. Pour comparer une courbe quelconque, obtenue par l'observation, avec la courbe théorique normale, M. Davenport (1) donne la méthode suivante : s'il s'agit d'une courbe ordinaire à ordonnées réunies par une ligne brisée ou par une courbe, la fréquence théorique pour chaque classe à l'écart $\frac{x}{\sigma}$ (x étant l'écart par rapport à la moyenne et σ étant l'écart type ou *standard deviation*) peut être prise directement dans la table insérée p. 672. Dans ce cas $\frac{x}{\sigma}$ est l'écart actuel d'après la moyenne exprimée en unités de la standard déviation et $\frac{y}{y_0}$ l'ordonnée correspondante, y_o étant pris comme égal à l'unité.

La méthode suivante est suivie pour obtenir l'entrée de la table :

Grouper en premier lieu les données de la manière indiquée pour trouver la moyenne par la méthode indirecte (Cfr. n° 229) c'est-à-dire en déterminant une origine arbitraire V*o* vers le milieu de la série et en donnant à chaque classe placée au-dessus une valeur négative croissante (1,

(1) C. B. Davenport, *Statistical Methods*, pp. 23 et 105.

Table de fréquence des erreurs sur 10000.

m	Nombre des erreurs	*m*	Nombre des erreurs	*m*	Nombre des erreurs	*m*	Nombre des erreurs
0	0	0.62	3242	1.34	6339	2.65	9261
0 01	54	0.64	3340	1.36	6410	2.70	9314
0.02	108	0 66	3438	1.38	6480	2.75	9364
0.03	161	0.68	3535	1.40	6550	2.80	9411
0.04	215	0 70	3632	1.42	6618	2.85	9454
0.05	269	0.72	3728	1.44	6686	2.90	9495
0.06	323	0.74	3823	1.46	6753	2 95	9534
0.07	377	0 76	3918	1.48	6818	3.00	9570
0.08	430	0.78	4012	1.50	6883	3.05	9603
0 09	484	0.80	4105	1.52	6947	3.10	9635
0.10	538	0.82	4198	1.54	7011	3 15	9664
0.12	645	0.84	4290	1.56	7073	3.20	9691
0.14	752	0.86	4381	1.58	7134	3.25	9716
0.16	859	0.88	4472	1.60	7195	3.30	9740
0.18	966	0.90	4562	1.64	7313	3.35	9762
0.20	1073	0.92	4651	1.68	7428	3.40	9782
0.22	1180	0.94	4739	1.72	7540	3.45	9800
0.24	1286	0.96	4827	1.76	7648	3.50	9818
0.26	1392	0.98	4914	1 80	7753	3.55	9834
0.28	1498	1.00	5000	1.84	7854	3.60	9848
0.30	1604	1.02	5085	1.88	7952	3.65	9862
0.32	1709	1.04	5170	1.92	8047	3.70	9870
0.34	1814	1.06	5254	1.96	8138	3.80	9896
0.36	1919	1.08	5337	2.00	8227	3.90	9915
0.38	2023	1.10	5419	2.05	8332	4.00	9930
0.40	2127	1.12	5500	2.10	8433	4.10	9943
0.42	2230	1.14	5581	2.15	8530	4.20	9954
0.44	2334	1.16	5660	2 20	8622	4 30	9963
0.46	2436	1.18	5739	2.25	8709	4.40	9970
0 48	2539	1.20	5817	2.30	8792	4.50	9976
0 50	2641	1.22	5894	2.35	8870	4.60	9981
0.52	2742	1.24	5971	2 40	8945	4.70	9985
0.54	2843	1.26	6046	2.45	9016	4.80	9988
0.56	2944	1.28	6121	2.50	9082	4.90	9991
0.58	3044	1.30	6194	2 55	9146	5.00	9993
0.60	3143	1.32	6267	2.60	9205		

2, 3, etc.) et à chaque classe située au-dessous de l'origine une valeur positive croissante d'après la même progression; la moyenne est égale au chiffre de la classe (V) augmenté de V_1; cette dernière valeur dont nous n'avons pas encore parlé est formée par une fraction dont le numérateur est la somme algébrique des produits $V - V_0$ par les fréquences, et le dénominateur est la somme des fréquences. Nous avons indiqué plus haut le procédé pour trouver la standard déviation (Cfr. n° 285). Il faut encore déterminer la fréquence y et la valeur y_0 : celle-ci est donnée par la formule

$$y_0 = \frac{n}{\sigma \sqrt{2\pi}}$$

et correspond à la fréquence maximum.

Eclaircissons la matière en l'illustrant d'un exemple numérique.

Reprenons les données utilisées à l'exemple 1 du chapitre II. (Livre II.)

Pimpinella Sax'fraga L

Classes	(f)	(ξ)	(f. ξ)	(f. ξ²)
5	1	— 6	— 6	36
6	5	— 5	— 25	125
7	9	— 4	— 36	144
8	22	— 3	— 66	198
9	38	— 2	— 76	152
10	62	— 1	— 62	62
11	61	0	— 271	
12	29	1	29	29
13	14	2	28	56
14	4	3	12	36
15	4	4	16	64
			+ 85	902

$$\Sigma\,(f.\,\xi) = -271 + 85 = -186$$

$$M - A = -\frac{186}{249} = -0.747 \quad M = 11 + (-0.747) = 10.253$$

$$\sqrt{\frac{902}{249}} = \sqrt{3.622} = \sigma\ 1.903$$

$$y_0 = \frac{n}{\sigma\sqrt{2\pi}} = \frac{249}{1.903 \times 2.506628} = \frac{249}{4.770113084} = 52.2$$

Le tableau pour la comparaison avec la courbe théorique se dispose comme suit :

Classes V	Écarts à M	$\frac{V - M}{\sigma}$	Valeurs d'après la table		y_0		y	f
5	— 6.253	3.28	0,0013	×	52.2	=	0,06786	1
6	— 5.253	2.76	0,0222	×	52.2	=	1,15884	5
7	— 4.253	2.23	0,0832	×	52.2	=	4,34304	9
8	— 3.253	1.71	0,2318	×	52.2	=	12,09996	22
9	— 2.253	1.18	0,4985	×	52.2	=	26,02170	38
10	— 1.253	0.65	0,8096	×	52.2	=	42,26112	62
11	253	0.13	0,9916	×	52.2	=	51,76152	61
12	1.747	0.92	0,6549	×	52.2	=	34,18578	29
13	2.747	1.44	0,3546	×	52.2	=	18,51012	14
14	3.747	1.97	0,1436	×	52.2	=	7,49592	4
15	4.747	2.49	0,0450	×	52.2	=	2,34900	4

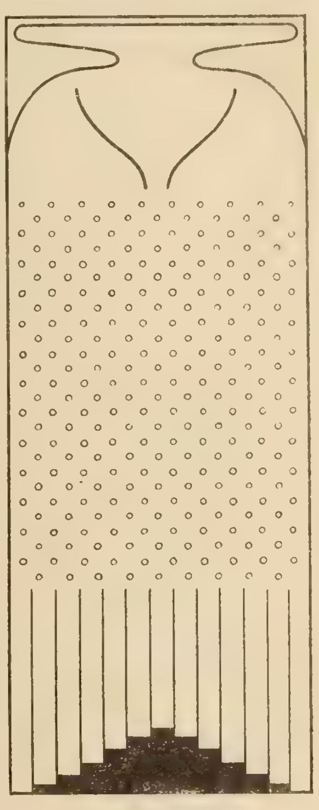
Les valeurs de $\frac{V - M}{\sigma}$ sont calculées dans la table suivante intitulée : table des ordonnées (x) de la courbe normale (1). Dans la quatrième colonne du tableau qui précède nous avons porté les valeurs lues dans la table ; dans la cinquième sont portées les valeurs des ordonnées de la courbe théorique et dans la sixième sont répétées les fréquences observées.

(1) C. B. Davenport, *Statistical methods*. New York, 1904, p. 118.

$\frac{x}{\sigma}$	0	1	2	3	4	5	6	7	8	9
0.0	100000	99995	99980	99955	99920	99875	99820	99755	99685	99596
0.1	99501	99396	99283	99158	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96420	96156	95882
0 3	95600	95309	95010	94702	94387	94055	93723	93382	93034	92677
0.4	92312	91939	91558	91169	90774	90371	89961	89543	89119	88688
0.5	88250	87805	87353	86896	86432	85962	85488	85006	84519	84060
0.6	83527	83023	82514	82010	81481	80957	80429	79896	79359	78817
0.7	78270	77721	77167	76610	76048	75484	74916	74342	73769	73193
0 8	72615	72033	71448	70861	70272	69681	69087	68493	67896	67298
0 9	66698	66097	65494	64891	64287	63683	63077	62472	61865	61259
1.0	60653	60047	51440	58834	58228	57623	57017	56414	55810	55209
1 1	54607	54007	53409	52812	52214	51620	51027	50437	49848	49260
1.2	48675	48092	47511	46933	46857	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38589	38058
1.4	37531	37007	36487	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31500	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22008	21627	21251	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1 9	16448	16137	15831	15530	15232	14939	14650	14364	14083	13806
2.0	13534	13265	13000	12740	12483	12230	11981	11737	11496	11259
2 1	11025	10795	10570	10347	10129	09914	09702	09495	09290	09090
2 2	08892	08698	08507	08320	08136	07956	07778	07604	07433	07265
2.3	07100	06939	06780	06624	06471	06321	06174	06029	05888	05750
2.4	05614	05481	05350	05222	05096	04973	04852	04734	04618	04505
2 5	04394	04285	04179	04074	03972	03873	03775	03680	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02684
2.7	02612	02542	02474	02408	02343	02280	02218	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01289	01252	01215	01179	01145
3.	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.	00000									

La méthode est différente si le matériel se compose d'une série de rectangles ou, ce qui revient au même, de données intermédiaires entre des limites de classes, comme c'est le cas dans les statistiques de salaires. On fait usage alors de tables spéciales dont on trouvera un spécimen dans l'ouvrage cité de C.-B. Davenport (1).

418. Galton a imaginé un appareil simple capable de reproduire mécaniquement la loi des erreurs; cet appareil a été remanié par le professeur Karl Pearson. Nous donnerons brièvement la description des deux appareils.


Fig. 44.

L'appareil de Galton consiste essentiellement en une boîte oblongue, dont la face supérieure est garnie d'une vitre; l'épaisseur de la boîte, en arrière de la vitre, est environ d'un quart de pouce; à la partie supérieure, deux pièces de bois sont disposées de manière à former entonnoir. Au-dessous se trouvent une quantité de lignes formées d'épingles placées à distance égale, en forme de quinconce, et plus bas treize compartiments constitués par des bandes verticales. On place dans ces compartiments une quantité de petits corps sphériques et on retourne l'appareil; toutes les « balles » se groupent dans la

(1) C. B. Davenport, *Statistical methods*, p. 119.

partie supérieure. On retourne ensuite la boîte pour la placer dans sa direction de travail.

Des bandes latérales dirigent les balles vers l'entonnoir et celles-ci tombent en cascade à travers le réseau d'épingles. Finalement, elles se casent dans les compartiments verticaux, en nombre inégal, mais de telle façon que leur groupement reproduit d'une manière sensiblement approchée la forme caractéristique de la courbe des erreurs. En effet, chaque balle en tombant rencontre une série d'obstacles, nombreux et indépendants, de manière que les accidents qui provoquent un écart vers la droite sont compensés par ceux qui causent une déviation vers la gauche; toutefois un grand nombre de balles en s'échappant de l'entonnoir suivent une direction verticale et la fréquence des balles qui s'échappent à droite et à gauche diminue plus rapidement que n'augmente la distance séparant les compartiments extrêmes de ceux du centre (1).

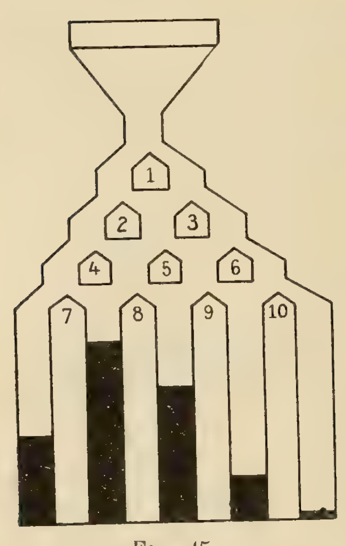


Fig. 45.

419. Le professeur Pearson a modifié ce dispositif de façon que les balles passant à travers les obstacles en forme de coin se trouvent dans le rapport $q^2 : pq : p^2$. Il a généralisé ensuite cette construction pour que les coins amovibles puissent être disposés de façon à donner tout rapport $q : p$.

Un entonnoir s'ouvre sur un espace d'un quart de pouce limité d'un côté par une feuille de bois et de l'autre par une glace. Au-dessous de l'entonnoir sont disposés des obstacles en forme de coin: immédiate-

(1) Galton, *Natural Inheritance*, ch. V., (Normal Variability), pp. 63-65.

ment en dessous se trouve un seul coin; à la tige suivante 2, à la troisième 3. Plus bas que les coins se trouvent des tiges de bois de même épaisseur et terminées en forme de coin; entre ces tiges, formant compartiments, les balles peuvent se rassembler. Le premier coin divise la masse des balles en deux parties, l'une se dirigeant vers la droite, l'autre vers la gauche de l'observateur.

Les coins 2 et 3 sont placés de façon à diviser les balles dans la même proportion, de façon que les balles passant à travers ces obstacles sont dans la proportion $q^2 : 2\ pq : p^2$ et plus loin, après la troisième ligne d'obstacles dans celle-ci : $q^3 : 3\ q^2p : 3\ q\ p^2 : p^3$, et finalement dans la suivante : $q^4 : 4q^3p : 6\ q^2\ p^2 : 4\ qp^3 : p^4$ (1).

IV. — Les types de distribution dérivés.

420. La courbe des erreurs accidentelles était connue depuis longtemps des mathématiciens et des physiciens, mais elle ne l'était que sous sa forme idéale, qui exige une distribution des chances toute spéciale. Les sciences firent un grand progrès le jour où l'on s'avisa que la loi des erreurs accidentelles pouvait être étendue à des phénomènes connus par l'observation; il est juste de rappeler que Quetelet fut un précurseur de cette méthode et que la voie qu'il avait ouverte conduisit à des résultats remarquables. L'hypothèse sur laquelle repose la conception actuelle de la loi des erreurs renferme une double assertion; à savoir : 1° que l'erreur dépend de causes extrêmement nombreuses; 2° que la fonction exprimant cette dépendance peut être traitée comme une fonction linéaire. La première de ces assertions paraît à la fois naturelle et exacte. « Dans toute observation, a écrit le Dr Glaisher, nous savons

(1) Cfr. Karl PEARSON : *Skew variation in homogeneous material*. Phil. Trans. Roy. Soc., série A, 1895, § 1er et Udny YULE : *An introduction to the theory of Statistics*, pp. 294-295.

que l'exactitude finale du résultat est altérée par de nombreuses causes indépendantes les unes des autres. L'état de santé de l'observateur, la perfection de ses sens, l'état de l'atmosphère, la construction plus ou moins parfaite des instruments qu'il emploie sont autant de causes qui contribuent à faire naître des erreurs. »

Il résulte de ceci que si l'on connaît la moyenne d'une série régulière et son écart probable, il n'est pas impossible de déterminer *a priori* le nombre des fréquences qui sont supérieures ou inférieures à cette moyenne, étant admis que leur nombre probable est égal de part et d'autre.

Cependant la loi normale des erreurs ne suffit pas à rendre compte de tous les phénomènes, car les recherches dirigées dans le domaine des sciences biologiques ont démontré que de nombreuses courbes sont déviées et présentent, par rapport à la courbe normale, d'importantes altérations. Les courbes, en dehors de la courbe normale, peuvent être plurimodales, ce qui indique alors que le matériel exprimé par la courbe n'est pas homogène, mais qu'il est composé de plusieurs types ayant leur individualité propre. Quetelet, sans avoir d'exemple à présenter pour justifier son hypothèse, avait déjà fait observer que si deux races distinctes se trouvaient comprises dans un même recensement, la courbe qui représenterait la hauteur des tailles, par exemple, aurait deux sommets qui annonceraient deux races différentes ayant des tailles moyennes inégales. Le Dr Bertillon, père, trouva un exemple remarquable de ces courbes plurimodales en analysant la taille des conscrits du département du Doubs; un sommet de la courbe correspond aux tailles comprises entre 1 m. 598 et 1 m. 652; l'autre aux tailles comprises entre 1 m. 679 et 1 m. 706 (1).

L'analyse de ces courbes plurimodales exige un effort

(1) Dr J. BERTILLON, *Cours élémentaire de statistique administrative*, Paris, 1895, p. 115.

mathématique considérable, soit que l'on décompose la courbe originale en deux courbes normales, soit que l'on procède à une double sériation conforme à la loi normale mais ayant chacune sa moyenne propre et son écart probable. Cette analyse nous entraînerait trop loin dans la voie des développements mathématiques.

Le cas de la déviation de la courbe des observations d'après une tendance qui agit tantôt dans un sens, tantôt dans un autre, est intéressant. La synthèse des cas de déviation pour les courbes unimodales a été faite par le Professeur K. Pearson dans une série de recherches d'un grand intérêt scientifique. Nous essayerons de faire saisir au lecteur la méthode d'après laquelle il y a lieu de procéder pour déterminer le type auquel appartient la courbe à analyser, et les règles à suivre pour comparer cette courbe à la courbe normale.

421. Les courbes unimodales simples se divisent en cinq types dont voici les formules :

I. $$y = y_0 \left(1 + \frac{x}{l_1}\right)^{m_1} \left(1 - \frac{x}{l_2}\right)^{m_2} \qquad (98)$$

Dans cette formule comme dans les suivantes, x désigne les abcisses, y_0 les ordonnées à leur origine, y la hauteur de l'ordonnée placée à la distance x à partir de y_0, l la partie de l'axe des abcisses exprimée en unités de classes, e la base du système népérien de logarithmes = 2.71828.

La courbe du type I est limitée dans les deux directions, c'est-à-dire que ses deux branches touchent l'axe des abcisses à une distance finie de l'ordonnée maximum, mais elle est asymétrique.

II. $$y = y_0 \left(1 - \frac{x^2}{l^2}\right)^{m} \qquad (99)$$

Cette courbe est analogue à celle du type I mais elle est symétrique dans les deux directions.

III. $$y = y_0 \left(1 + \frac{x}{l}\right)^p e^{-x/d} \tag{100}$$

Le type III exprime les courbes qui ne sont limitées que dans une seule direction, c'est-à-dire qu'une de leurs branches atteint l'axe des abcisses à une distance finie de l'ordonnée maximum, tandis que l'autre se rapproche indéfiniment de cet axe sans jamais l'atteindre ; par l'opposition de la direction des branches qui les composent, ces courbes sont évidemment asymétriques.

IV. $$y = y_0 \cos \theta^{2m} e^{-\tau\theta}, \text{ ou tangente } \theta = \frac{x}{l} \tag{101}$$

Courbe non limitée dans les deux directions, comme la courbe normale des erreurs, mais asymétrique. Les branches de la courbe se rapprochent donc de l'axe des abcisses mais sans l'atteindre jamais.

V. $$y = y_0 x^{-p} e^{-\gamma/x} \tag{102}$$

Courbe analogue à celle du type IV, mais symétrique.

422. Le problème soulevé par l'analyse des courbes de fréquence se ramène à trois points : il faut, en premier lieu, trouver le type auquel appartient une courbe quelconque ; ensuite, il faut déterminer le tracé de la courbe au moyen de formules appropriées à chaque cas ; enfin, il y a lieu de comparer les résultats de l'observation à la courbe théorique. De ces trois aspects, le premier est le seul qui puisse être exposé à des lecteurs qui ne sont pas familiers avec les mathématiques supérieures : c'est le seul que nous aborderons ici, en renvoyant pour de plus amples explications aux ouvrages spéciaux, la matière sortant des limites d'un traité.

Pour déterminer le type auquel se rattache une courbe, commençons par chercher la moyenne des grandeurs ou

variables : adoptons une classe proche de la moyenne supposée et appelons-la V_o. Cette classe servira d'origine et portera en regard le chiffre 0. Les classes situées dans la partie supérieure sont numérotées de 1 à n, en partant de zéro : ce sont les classes dont les variables sont considérées comme négatives. Celles qui sont situées dans la partie inférieure sont de même numérotées de 1 à n, en partant de zéro : ce sont les classes à variables positives. Les chiffres des classes sont inscrits dans la seconde colonne intitulée $V - V_o$, V étant l'indication de la classe (première colonne). Dans la troisième colonne nous transcrivons les fréquences dans l'ordre de leurs classes respectives et nous en faisons le total. Nous multiplions les fréquences par $V - V_o$ et nous inscrivons les produits dans une quatrième colonne intitulée $f\,(V - V_o)$; nous faisons la somme algébrique des quantités positives et négatives. La cinquième colonne est formée des carrés de $V - V_o$ multipliant les fréquences de la classe correspondante, ou bien encore en multipliant les premiers produits obtenus par $V - V_o$. La somme algébrique de ces produits est de nouveau faite, puis divisée par n ; le quotient est appelé ν_2, d'après la formule :

$$\nu_2 = \frac{\Sigma\,(V - V_o)^2}{n} \qquad (103)$$

On procède de même pour les troisième et quatrième puissances des nombres considérés. Les formules :

$$\nu_3 = \frac{\Sigma(V - V_o)^3}{n} \quad (104) \qquad \nu_4 = \frac{\Sigma(V - V_o)^4}{n} \quad (105)$$

expriment les mêmes opérations que dans les formules précédentes, sauf la puissance à laquelle les nombres sont élevés.

Les formules ci-dessus sont appliquées dans l'exemple suivant à la série du *Pimpinella Saxifraga L* que nous avons déjà utilisée à plusieurs reprises.

Pimpinella saxifraga L. (1)

1	2	3	4	5	6	7
V	$V - V_o$	f	$f(V - V_o)$	$f(V - V_o)^2$	$f(V - V_o)^3$	$f(V - V_o)^4$
5	— 5	1	— 5	25	— 125	625
6	— 4	5	— 20	80	— 320	1280
7	— 3	9	— 27	81	— 243	729
8	— 2	22	— 44	88	— 176	352
9	— 1	38	— 38	38	— 38	38
10	0	62	— 134	—	— 902	—
11	1	61	61	61	61	61
12	2	29	58	116	232	464
13	3	14	42	126	378	1134
14	4	4	16	64	256	1024
15	5	4	20	100	500	2500
			197	779	1427	8207
			+ 63		+ 525	

$$\Sigma\,(V - V_o) = -\ 134 + 197 = +\ 63$$

$$\Sigma\,(V - V_o)^2 = 779$$

$$\Sigma\,(V - V_o)^3 = -\ 902 + 1427 = +\ 525$$

$$\Sigma\,(V - V_o)^4 = 8207.$$

Les totaux des colonnes 4 à 7 sont divisés par le nombre des observations et les quotients obtenus sont les moments autour de 10.

---

(1) Cfr. W. Palin ELDERTON, *Frequency-curves and correlation*, p. 16. C. B. DAVENPORT : *Statistical methods.*

D'après les formules :

$$\frac{\Sigma (V - V_o)}{n} = \frac{+ 63}{249} = + .253$$

$$\frac{\Sigma (V - V_o)^2}{n} = \frac{779}{249} = 3.128$$

$$\frac{\Sigma (V - V_o)^3}{n} = \frac{+ 525}{249} = 2.108$$

$$\frac{\Sigma (V - V_o)^4}{n} = \frac{8207}{249} = 32.959$$

Les expressions ci-dessus sont appelées « moments » de la courbe.

Il faut maintenant déterminer la place des moments autour de la moyenne.

Deux méthodes sont applicables. La première est celle à employer quand il s'agit de variables qui, d'après leur nature, sont exprimées en nombres entiers, par comptage. C'est le cas des caractères de variabilité du *Pimpinella saxifraga L.* que nous avons pris pour exemple. Les formules applicables à cette recherche sont les suivantes :

$$\mu_1 = 0 \qquad \textbf{(106-111)}$$

$$\mu_2 = \nu_2 - \nu_1^2$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$$

$$\mu_5 = \nu_5 - 5\nu_1\nu_4 + 10\nu_1^2\nu_3 - 10\nu_1^3\nu_2 + 4\nu_1^5$$

$$\mu_6 = \nu_6 - 6\nu_1\nu_5 + 15\nu_1^2\nu_4 - 20\nu_1^3\nu_3 + 15\nu_1^4\nu_2 - 5\nu_1^6.$$

Ces formules s'écrivent également en remplaçant ν_1 par d; lorsque le calcul des moments se fait autour de la verticale centrale, comme dans le tableau ci-dessus, la valeur de d est la même que ν_1 (1).

(1) Cfr. W. Palin ELDERTON, *Frequency-curves and correlation*, p. 18.

Appliquons les formules :

$$\mu_2 = 3.128 - (+.253)^2 = 3.128 - 0.064009 = 3.063991$$

$$\mu_3 = 2.108 - 3(.253 \times 3.128) + 2(.253)^3$$
$$= 2.108 - 2.374152 + .32388554 = .057733$$

$$\mu_4 = 32.959 - 4(.253 \times 2.108) + 6(.064009 \times 3.128)$$
$$- 3(.04097152081) =$$
$$32.959 - 4(.533324) + 6(.200220152)$$
$$- 3(.04097152081) =$$
$$32.959 - 2.133296 + 1.201320912 - .12291456243$$
$$= 31.9041102877.$$

Le travail arithmétique ci-dessus peut être remplacé par l'usage des logarithmes.

Davenport en a donné un exemple détaillé auquel nous renvoyons le lecteur (1).

423. La méthode qui vient d'être indiquée est à la fois la plus directe et la mieux appropriée pour déterminer les moments; toutefois, il en est une autre qui a été indiquée et employée par un actuaire anglais, M. G. F. Hardy (2). Le mode de calcul usité dans cette méthode est le même que celui des « pourcentages accumulés » dont on se sert fréquemment dans les statistiques de salaires (3). Le premier moment est connu en additionnant les « pourcentages accumulés » des fréquences ($\Sigma f(n) = V_o$), et le second moment en additionnant de même les pourcentages du premier moment.

Dans la notation spéciale usitée, S_2 désigne le second moment; S_3 donne chaque fonction multipliée par des coefficients de la forme $\frac{n(n+1)}{2}$ ou $\frac{n^2+n}{2}$:

$$S_3 = f(1) + 3f(2) + 6f(2) + \ldots\ldots\ldots + \frac{n.\overline{n+1}}{2} f(n) \qquad \mathbf{(112)}$$

(1) Cfr. B. DAVENPORT, *Statistical methods*, p. 36.

(2) Référence à M. G. F. HARDY, dans W. Palin Elderton, *Frequency curves and correlation*, p. 19.

(3) Voyez sur cette méthode : A. JULIN, *Précis du cours de statistique générale et appliquée*, 3e édition, Bruxelles, Misch et Thron, 1912, p. 71.

ce qui donne $\frac{\nu_2 + \nu_1}{2}$ expression dans laquelle ν est écrit pour le moment parce que par définition le t^{me} moment (ν_t) de la distribution entière est donné par la somme de $n^t f(n)$ pour toutes les valeurs de n. Enfin S_4 et S_5 donnent chacun une fonction multipliée par $\frac{n^3 + 3n^2 + 2n}{6}$ et $\frac{n^4 + 6n^3 + 11n^2 + 6n}{24}$ respectivement (1).

Nous avons donc les opérations ci-après :

$$S_2 = \nu_1$$
$$S_3 = \tfrac{1}{2}(\nu_2 + \nu_1)$$
$$S_4 = \tfrac{1}{6}(\nu_3 + 3\nu_2 + 2\nu_1)$$
$$S_5 = \tfrac{1}{24}(\nu_4 + 6\nu_3 + 11\nu_2 + 6\nu_1)$$

Telles sont les équations qui permettent de calculer les moments autour d'une origine choisie, mais s'il convient de rechercher les moments autour de la moyenne, les expressions ci-après sont préférables :

$$\nu_2 = 2\,S_3 - d\,(1 + d)$$
$$\nu_3 = 6\,S_4 - 3\,\nu_2\,(1 + d) - d\,(1 + d)\,(2 + d)$$
$$\nu_4 = 24\,S_5 - 2\,\nu_3 \left\{ 2\,(1 + d) + 1 \right\} - \nu_2 \left\{ 6\,(1 + d)\,(2 + d) - 1 \right\} - d\,(1 + d)\,(2 + d)\,(3 + d)$$

424. Lorsqu'il s'agit de séries disposées de façon à calculer la moyenne par le procédé ordinaire, la suite des calculs, dans la méthode de M. Hardy, se dispose de la manière suivante : 1° les nombres bruts de la série sont rapportés à leur somme, prise pour l'unité; il est recommandé de pousser les calculs au moins jusqu'à la troisième décimale; les nombres proportionnels rangés dans leur ordre forment la première colonne du tableau; 2° la seconde co-

(1) Cfr. W. Palin Elderton, *loc. cit.*, p. 21.

lonne forme le résultat, S_2; le dernier nombre de la première colonne est inscrit au bas de la seconde; on y ajoute l'avant-dernier de la première colonne et on inscrit ce résultat dans la seconde immédiatement au-dessus du nombre inscrit précédemment, et ainsi de suite; le dernier nombre inscrit au-dessus de la colonne doit former l'unité suivie de ses parties décimales; enfin, on fait la somme de tous les nombres de la colonne : cette somme $= S_2$; ajoutée à une valeur arbitraire prise comme origine, et multipliée par l'intervalle de classe elle forme la moyenne; 3° on procède de même pour la troisième, la quatrième, la cinquième colonne en utilisant toujours les données inscrites dans la colonne précédente; les sommes obtenues correspondent à S_3, S_4, S_5, respectivement.

Le procédé est différent si l'on emploie la méthode indirecte pour trouver la moyenne. Dans ce système, certaines quantités sont négatives, d'autres positives : les moments, pour les termes appartenant à la partie négative, sont formés par multiplication des pouvoirs des quantités négatives. Si n est une valeur négative, la formule $\Sigma\, nf\,(n)$ devient $\Sigma - nf\,(-n)$ et la formule

$$\Sigma \frac{n(n+1)}{2} f(n) \quad \text{se change en} \quad \Sigma \frac{-n(-n+1)}{2} f(-n)$$

ou

$$\Sigma \frac{n(n-1)}{2} f(-n); \tag{113}$$

de même nous avons :

$$\Sigma \frac{-n(-n+1)(-n+2)}{6} f(-n)$$

ou

$$\Sigma \frac{n(n-1)(n-2)}{6} f(-n); \tag{114}$$

enfin nous avons :

$$\Sigma \frac{-n(-n+1)(-n+2)(-n+3)}{24} f(-n)$$

ou

$$\Sigma \frac{n(n-1)(n-2)(n-3)}{24} f(-n) \tag{115}$$

L'exemple suivant reprend les données utilisées précédemment au n° 422.

Fréquences	Première somme	Deuxième somme	Troisième somme	Quatrième somme	Cinquième somme	Sixième somme
4	4	4	4	4	4	4
20	24	28	32	36	40	
36	60	88	120	156		
89	149	237	357			
152	301	538				
249						
245	450	791	1332	2153	3350	5035
117	205	341	541	821	1197	
56	88	136	200	280	376	
16	32	48	64	80	96	
16	16	16	16	16	16	

$$S_2 = .791 - .538 = .253$$

$$S_3 = 1.332 + .357 = 1.689$$

$$S_4 = 2.153 - .156 = 1.997$$

$$S_5 = 3.350 + .040 = 3.390$$

$$S_6 = 5.035 - .004 = 5.031$$

Or $$\nu_2 = 2S_3 - d\,(1 + d),$$

soit dans l'exemple numérique

$$2 \times 1.689 - .253 \times 1.253 = 3.063991 \text{ etc.}$$

425. Pearson a démontré que la classification de tout polygone de fréquence obtenu par l'observation dé-

pend de la valeur de ce qu'il appelle sa « fonction critique » F (*critical function*). Or

$$F = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} \qquad (116)$$

Or,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}; \qquad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

Déterminons ces valeurs :

$$\mu_3^2 = (.057733)^2 = .003333099289$$

$$\mu_2^3 = (3.063991)^3 = 28.764873$$

$$\mu_4 = 31.904110$$

$$\mu_2^2 = (3.063991)^2 = 9.388041$$

$$\beta_1 = \frac{.003333}{28.764873}; \qquad \beta_2 = \frac{31.904110}{9.388041}$$

d'où

$$\beta_1 = .000115870; \qquad \beta_2 = 3.398377$$

et

$$F = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}$$

ou

$$F = \frac{.000115870 \times (40.939253)}{4(13.593161390)(.796406390)}$$

$$= \frac{.004743631245}{41.302699917464} = .001148503$$

Or, si F > 0 et < 1, on a une courbe du type IV, ce qui est ici le cas.

La courbe du *Pimpinella Saxifraga L.* est donc du type IV, qui est le type habituel des courbes des phénomènes biologiques.

426. Si l'on veut comparer le polygone observé avec le polygone théorique afin de déterminer le degré d'exactitude avec lequel le premier suit la courbe du second, on fera la différence entre la valeur théorique des fréquences de chaque classe et sa valeur observée; on porte ensuite ces différences au carré et on divise leur somme par la valeur théorique : la racine carrée du quotient constitue l'indice d'exactitude cherché. La formule

$$\sqrt{\Sigma \frac{\delta^2}{y}} \tag{117}$$

dans laquelle δ représente les différences et y la valeur théorique donne la valeur du coefficient, qui est souvent désigné par la lettre Δ (K. Pearson.)

427. Lorsqu'il s'agit de variables dont les classes sont exprimées par des chiffres limites, par exemple des salaires compris entre telle somme et telle autre, les moments de la courbe se déterminent à l'aide d'une formule différente de la première.

Soient connues les valeurs de ν_1, ν_2, ν_3, ν_4, les valeurs de μ s'obtiennent au moyen des formules suivantes :

$$\mu'_1 = 0 \tag{118}$$

$$\mu'_2 = \left[\nu_2 - \nu_1^2 - \frac{1}{12} \right] \lambda^2 ; \tag{119}$$

$$\mu'_3 = \left[\nu_3 - 3 \nu_1 \nu_2 + 2 \nu_1^3 \right] \lambda^3 ; \tag{120}$$

$$\mu'_4 = \left[\nu_4 - 4 \nu_1 \nu_3 + 6 \nu_1^2 \nu_2 - 3 \nu_1^4 - \frac{1}{2} (\nu_2 - \nu_1^2) + \frac{7}{240} \right] \lambda^4 ; \tag{121}$$

$$\mu'_5 = \left[\nu_5 - 5 \nu_1 \nu_4 + 10 \nu_1^2 \nu_3 - 10 \nu_1^3 \nu_2 + 4 \nu_1^5 - \frac{5}{6} \mu_3 \right] \lambda_5 ; \tag{122}$$

dans ces expressions, λ est l'intervalle de classe exprimé dans l'unité de la moyenne.

428. Les caractères distinctifs des courbes de Pearson se révèlent au moyen de la valeur de la fonction critique F.

Pour la courbe	I,	nous avons	$F < 0$
»	II,	»	$F = 0$, $\beta_1 = 0$, β_2 n'égale pas 3 ;
»	III,	»	$F = \infty$
»	IV,	»	$F > 0$ et < 1
»	V,	»	$F = 1$.

V. — Quelques déterminations d'erreurs probables.

429. Les formules suivantes permettront de rechercher l'erreur probable de quelques mesures importantes employées en statistique.

Erreur probable de la moyenne. — La mesure à laquelle le professeur Pearson a donné le nom de « Standard déviation » dérive évidemment de la théorie des erreurs d'observation : c'est à raison de ce caractère qu'elle a été nommée « erreur moyenne » par Gauss, « error of mean square » par Airy et quelquefois « mean square error ». Nous rappelons sa formule

$$\sigma = \sqrt{\frac{1}{N} \Sigma (x^2)} \qquad (28)$$

La « Standard déviation » est un des éléments principaux de la formule de l'erreur probable de la moyenne : celle-ci est en effet connue si l'on multiplie la constante 0.6745 par la « Standard déviation » divisée elle-même par la racine carrée du nombre de variables. La formule, en effet, est la suivante :

$$\pm 0.6745 \frac{\sigma}{\sqrt{n}} \qquad (123)$$

Appliquons la formule à un exemple numérique :

Dans l'observation de la « pimpinella Saxifraga L. » la Standard déviation $\sigma = 1.461$. Le nombre $n = 249$; $\sqrt{249} = 15.7797338$. De là, on a :

$$\pm 0.6745. \frac{1.461}{15.780} = 0,06239125.$$

L'erreur moyenne est d'autant plus petite que le nombre des variables est plus grand; cependant son exactitude n'augmente pas comme le nombre des variables, mais seulement dans la proportion de la racine carrée de ce nombre.

430. *Erreur probable de la médiane.* — La Standard déviation d'un percentile correspondant à une proportion p ( ) est connue par la formule

$$\sigma x_p = \frac{\sigma}{y_p} \sqrt{\frac{p\,q}{n}} \qquad (124)$$

Dans le cas d'une distribution analogue à celle de la courbe normale, les valeurs de y_p sont données directement par des tables dressées par M. Sheppard, d'après lesquelles cette valeur est pour

la médiane = 0.3989423

Après l'insertion de la constante 0.3989423 dans l'équation ci-dessus, on tire la constante 0.84535.

De là, la formule

$$\pm\, 0.84535 \frac{\sigma}{\sqrt{n}}$$

431. *Erreur probable de la Standard déviation.* — L'expression générale de l'erreur probable de la Standard déviation dans le cas d'une distribution de n'importe quelle forme est :

$$= \sqrt{\frac{\mu_4 - \mu_2^2}{4\mu_2 . n}}\ (1) \qquad (125)$$

Mais dans le cas d'une distribution normale, la formule est la suivante :

$$\pm\, 0.6745 \frac{\sigma}{\sqrt{2n}} \qquad (126)$$

(1) G. U. YULE, *An introduction to the theory of statistics*, p. 347.

Appliquons la seconde formule à l'exemple numérique calculé au n° 422. Nous avons :

$$\sigma = 52.2 \qquad n = 2.49 \qquad \sqrt{2n} = 22.3159136$$

$$\pm\, 0.6745 \frac{\sigma}{\sqrt{2n}} = 1.5776555$$

432. *Erreur probable du coefficient de corrélation ou de covariation r* (1).

Pour trouver l'erreur probable du coefficient de corrélation, il faut porter au carré le coefficient de corrélation et le soustraire de l'unité puis diviser le reste par la racine carrée du nombre de variables, le tout étant multiplié par la constante 0.6745. La formule s'écrit :

$$0.6745 \frac{(1 - r^2)}{\sqrt{n}} \qquad (127)$$

Le coefficient de corrélation entre les indices de la production et ceux des échanges en Belgique, entre les années 1880 et 1908 est 0.982.

Nous avons

$$r^2 = 0.964324$$

$$1 - r^2 = 0.035676$$

$$n = 29\,; \quad \sqrt{29} = 5.3851648$$

$$0.6745 \cdot \frac{(1 - r^2)}{\sqrt{29}} = \frac{0.035676}{5.3851648} = 0.6745 \times 0.006604 = 0.0044543980$$

433. *Erreur probable du coefficient de régression.* — Pour trouver l'erreur probable du coefficient de régression on a recours à la formule :

$$0.6745 \,.\, \frac{\sigma_1 \sqrt{1 - r_{12}^2}}{\sigma_2 \sqrt{n}} \qquad (128)$$

dans laquelle la dernière expression peut être remplacée par celle-ci :

$$\frac{\sigma_{1.2}}{\sigma_2 \sqrt{n}} \qquad (129)$$

(1) Cfr. W. Palin Elderton, *Frequency-curves and correlation*, pp. 136-138.

434. *Références.*

BENINI (R.). *Principii de Statistica metodologica*, Turin, 1906.

BOREL (E.). *Eléments de la théorie des probabilités*, Paris, Hermann et fils, 1909.

BOSCO (A.). *Lezioni di statistica*, Roma, 1905.

BOUDIN (E. J.). *Leçons de calcul des probabilités faites à l'Université de Gand de 1846 à 1890*, par E. J. Boudin, publiées par Paul Mansion, Paris, Gauthier-Villars, 1916.

BOWLEY (A. L.). *Elements of Statistics*, seconde édition, London, King and Son, 1902.

CARVALLO (E.). *Le calcul des probabilités et ses applications*, Paris, Gauthier-Villars, 1912.

DAVENPORT (C. B.). *Statistical methods*, seconde édition, New York, Wiley and Sons, 1904.

EDGEWORTH (F. Y.). The Law of Error, Cambridge Phil. Trans., vol. XX, 1904; « The generalised law of Error, of law of great Numbers », *Journ. Roy. Stat. Soc.*, vol. LXIX 1906; « On the Representation of statistical Frequency by a Curve », *Journ. Roy. Stat. Soc.*, vol. LXX, 1907.

ELDERTON (W. Palin). *Frequency-curves and correlation*, London, Layton. 1906.

GALTON (Francis). *Natural inheritance*, London, Macmillan, 1889.

JEVONS (W. Stanley). *The Principles of Science*, London, Macmillan 1892.

MANSION (P.). Voyez *supra* BOUDIN.

MONTESSUS (de) *Leçons élémentaires sur le calcul des probabilités*, Paris, Gauthier-Villars, 1910.

PEARSON (K.). *Skew variation in homogeneous material*, Phil. Trans. Roy. Soc. Series A, vol. Cl. XXXVI, 1895.

QUETELET (A.). *Lettres sur la théorie des probabilités*, Bruxelles, Hayez, 1846.

VENN (J.). *The logic of chance*, London, Macmillan, 1888.

WRIGHT (T. Wallace). *The adjustement of observations*, 2e édition, Londres et New-York, 1906.

YULE (G. Udny). *An introduction to the theory of statistics*, London, Griffin, 1911.

TABLE ANALYTIQUE DES MATIÈRES.

Chapitre III. — Caractères propres à la statistique.

Chapitre IV. — Division de la matière.

LIVRE PREMIER

Technique du relevé statistique.

PREMIÈRE SECTION. — Le relevé statistique ou relevé direct.

CHAPITRE PREMIER. — Généralités, définition, division.

CHAPITRE II. — Organisation du relevé statistique.

II. *Influence de la centralisation sur l'exactitude des résultats.* — Les deux systèmes en présence. Absence de critique dans le système décentralisé ; tout repose sur l'agent recenseur (106).

III. *Nécessité de la critique statistique.* — Le chiffre fascinateur (107) ; l'examen du document est indispensable en statistique comme en histoire (108) ; première notion des erreurs constantes et accidentelles (109) ; division de la matière (110).

Chapitre II. — Critique de sincérité

Mobiles psychologiques influençant la sincérité statistique — Distinction entre erreurs constantes et erreurs fortuites (111) ; influence de la crainte ; crainte des mesures fiscales et de réglementation (112) ; moyens pour atténuer cette cause d'erreurs (113) ; mesures spéciales pour corriger des erreurs constantes (114) ; craintes de désavantages d'ordre moral (115) ; paresse, négligence, mauvaise volonté des recensés et des agents recenseurs (116) ; le statisticien et sa responsabilité scientifique (117).

Chapitre III. — Critique d'exactitude.

En quoi consiste la critique d'exactitude (118).

I. *Vérification interne.* — Recherche des lacunes de la statistique ; moyens de les constater (119) ; utilité des comparaisons (120) ; recherche des multiples emplois ; exemples (121) ; recherche des erreurs involontaires, contradictions, etc. ; l'auteur du document a-t-il répondu ? (122) ; l'auteur du document a-t-il compris ? (123) ; rôle des questions de contrôle (124) ; procédés de revision ; exemples (125).

II. *Vérification externe* (126).

Chapitre IV. — Précision des résultats.

Règle de la recherche de la précision (*form. 4*) (127) ; difficulté résultant de l'absence, dans les recherches statistiques, de mesures types (128) ; exemples de comparaison à des mesures plus exactes (129) ; pourquoi ces cas sont rares (130) ; comparaison entre plusieurs mesures (131) ; effet de plusieurs erreurs dans les additions et les moyennes (*form 5*) (132) ; simplification des nombres cités ; règles (133) ; références sur la critique statistique (134).

Section IV.

Le dépouillement et la présentation des données statistiques.

Chapitre premier. — Préparation du relevé.

I. *Généralités et division de la matière.* — Processus du dépouillement (135); le dépouillement prépare le travail scientifique statistique (136); à quelles données s'étend le dépouillement (137).

II. *Conditions générales du dépouillement.* — Il ne peut s'écarter trop du matériel original (138); le cadre statistique (139); exemples (140); formation des classes dans un groupement statistique (141); exemple numérique (142); formules (143) (*form 6-7*); utilisation dans le calcul des probabilités (144); exemples de tableaux de complexité croissante (145); nombre de combinaisons possibles entre données statistiques (146).

III. *Des classifications statistiques.* — Importance des classifications (147); chaque espèce de statistique a son principe de classification (148); inconvénients des classifications trop brèves ou trop étendues (149); danger d'utiliser certains principes de la classification générale des sciences (150); des nomenclatures alphabétiques (151); classifications homogènes et hétérogènes (152); du groupement des faits statistiques dans les cadres (153); précision des divisions (154); point de départ des divisions (155).

IV. *Préparation des tableaux.* — Relations logiques entre les divisions des tableaux à double entrée; exemple (156).

Chapitre II. — Exécution du dépouillement.

I. *Organisation du dépouillement.* — Méthodes générales : décentralisation, centralisation (157); avantages de la méthode centralisatrice (158); objections faites à cette méthode (159).

II. *Méthodes de dépouillement.* — Méthode de dépouillement par pointage (160); ses désavantages (161); dépouillement par fiches (162); utilisation de ce procédé (163); modèles de fiches (164); avantages des fiches (165).

Chapitre III. — Le dépouillement statistique et le calcul par les machines.

I. *Machines à dépouiller.* — Nécessité dans les grandes opérations de simplifier le travail (166); machine électrique de M. Hollerith; la fiche (167); le Key-Punch (*fig. 1*) (168); le Gang-Punch (*fig. 2*) (169); la

machine à classer (*fig. 3*) (170); la machine à dépouiller (*fig. 4, 5, 6*) (171); perfectionnements nouveaux (172); classi-compteur imprimeur de M. March (*fig. 7*) (173).

II. ***Machines à calculer.*** — Premiers essais (174); arithmographe Tronchet (*fig. 8*) (175); additionneur Roth (*fig. 9*) (176); machines à additionner, à touches (*fig. 10 à 12*) (177); multiplication par additions successives, soustraction par l'emploi des nombres complémentaires (178); machines à multiplier (*fig. 13*) (179); rouleau diviseur (*fig. 14*) (180).

III. ***Avantages des machines*** (181).

Chapitre IV. — **La présentation des résultats statistiques.**

I. ***Règles pratiques de la présentation statistique*** (182); remarques sur la forme matérielle de la présentation (183); références sur le dépouillement et la présentation (184).

LIVRE SECOND

Procédés d'analyse du matériel statistique.

Généralités et division de la matière.

Le travail de mise en œuvre et d'analyse succède au récolement et au classement (185). La classification des opérations se base sur leur ordre logique et non sur leur caractère mathématique (186).

Chapitre premier. — **Séries, sériation, distribution.**

I. *Les séries statistiques.* — Une succession de données forme une série: les phénomènes sont observés selon les divisions du temps, de l'espace ou en eux-mêmes; classement de séries (187). Séries basées sur une mesure du temps : l'année, le semestre, le mois; exemples (188); les divisions peuvent être plus courtes : la journée, l'heure; définition de l'échelle de la série; exemples (189-190). Séries basées sur les divisions de l'espace; ce classement n'échappe pas à tout arbitraire, mais il a aussi ses avantages; exemple (191). Etude des distributions, troisième modalité des séries, exemple (192). Autre

Chapitre III. — La dispersion et ses mesures.

caractéristiques (359); la statistique graphique est utile, mais il est fréquent de tomber dans l'excès; cas dans lesquels son usage est abusif (360); *B. Diagrammes orthogonaux*; diagrammes de succession au historigramme; leur utilité (361); application; transformation des données numériques en un historigramme (*fig. 38*); avantages de la figure sur les nombres (362); rapport entre l'espacement sur l'axe des abcisses et l'axe des ordonnées; règles générales (363); procédé de M. J. Bertillon (364); convention proposée par l'Institut international de statistique (*fig. 39*, (365); second exemple (366); influence de la position de la moyenne sur l'allure de la courbe (367); définition du diagramme de distribution; différences entre les deux espèces de diagrammes (*form. 80*) (368); le polygone de fréquence et l'histogramme (369); les termes extrêmes peuvent-ils être négligés? règle du calcul, application (370); application aux courbes de fréquence des conventions admises en ce qui concerne les courbes chronologiques (371); *C. Courbes logarithmiques*; inconvénients, au point de vue de la précision, des diagrammes orthogonaux; usage des courbes logarithmiques (372); particularités des courbes logarithmiques (373); cas où leur utilité spéciale apparaît surtout; application *fig. 40*); construction et interprétation de la courbe (374); courbes à double échelle logarithmique (375); *D. Diagrammes polaires*; fondement géométrique, construction, exemple (376); *E Cartogrammes*; définition, étendue de la définition (377); cartes avec diagrammes, leur usage (378); application spéciale de ce procédé (379); cartes teintées; définition; nombre de divisions géographiques; exposé du procédé; impression fausse à éviter (380); observations sur le nombre de teintes à employer (381); cartes à bandes; exposé du procédé (382); cartes avec niveau; en quoi elles consistent; leurs inconvénients (383); autres figures coloriées non accompagnees de cartes; procédés de MM. Foville et Benini (384); stéréogrammes; différences avec les diagrammes; pouvoir étendu de représentation; principes de leur construction; application (385); les stéréogrammes sont d'un usage rare et peu pratique (386).

III. *La statistique graphique comparative.* — Rôle des représentations graphiques sous le rapport comparatif (387); règles et procédés de comparabilité (388); réunion des données proportionnelles et absolues (389); comparaison des courbes de distribution; nécessité de l'usage des nombres proportionnels; comment utiliser en même temps les données absolues; procédé de M. March (*fig. 41*) (390); comparaisons à l'aide de cartogrammes; procédés de M. Cheysson (391).

IV. *La statistique graphique comme instrument d'investigation.* — Méthode graphique des pourcentiles de Sir F. Galton; manière de tracer la courbe de Galton; application; interprétation (*fig. 42*) (592); méthode graphique pour la recherche de la médiane de M. Yule;

LIVRE III

La loi des erreurs.

I — Définitions et généralités.

II. — Notions sur les probabilités.

III. — La loi des erreurs.

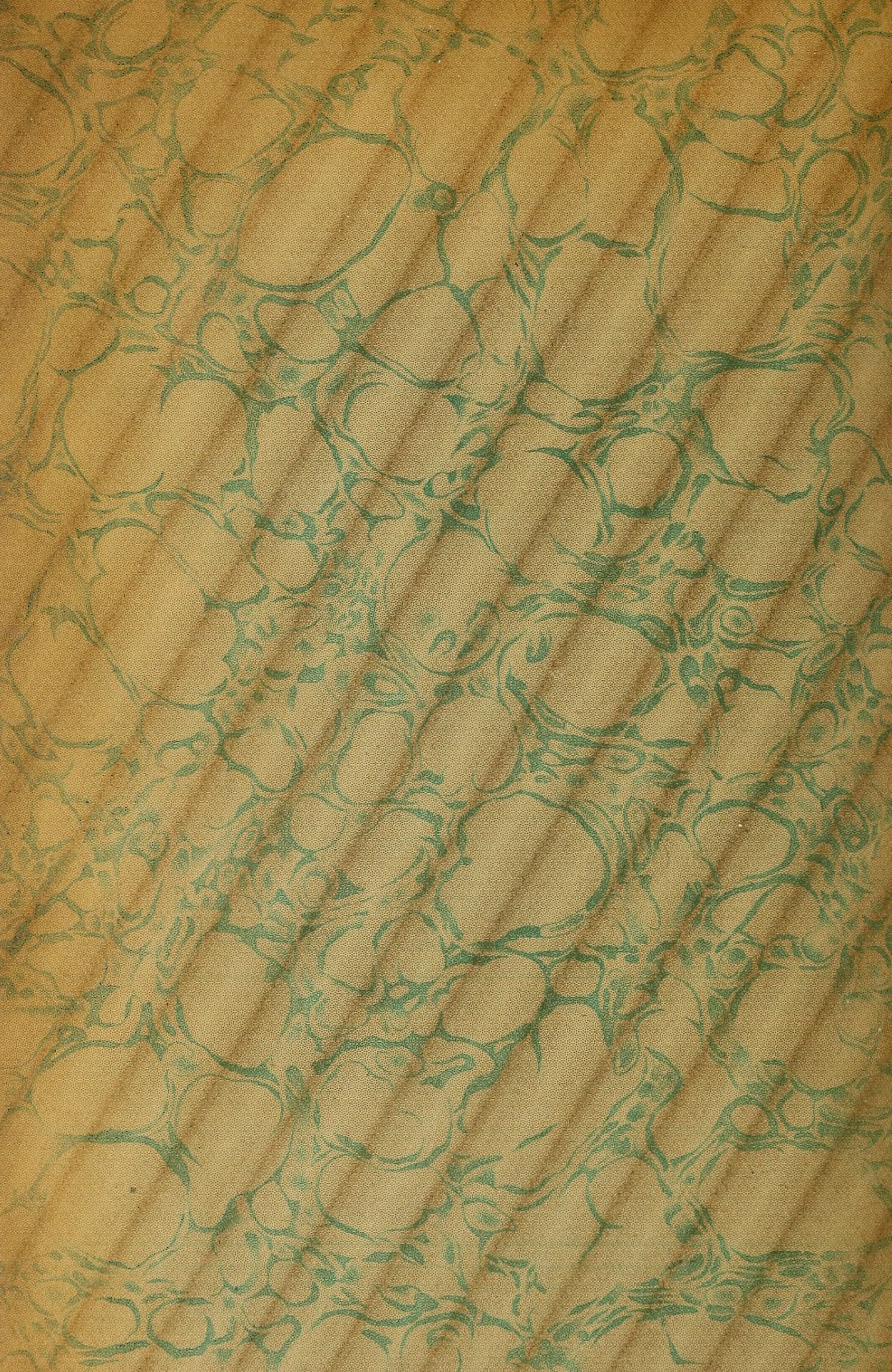
IV. — Les types de distribution dérivée.

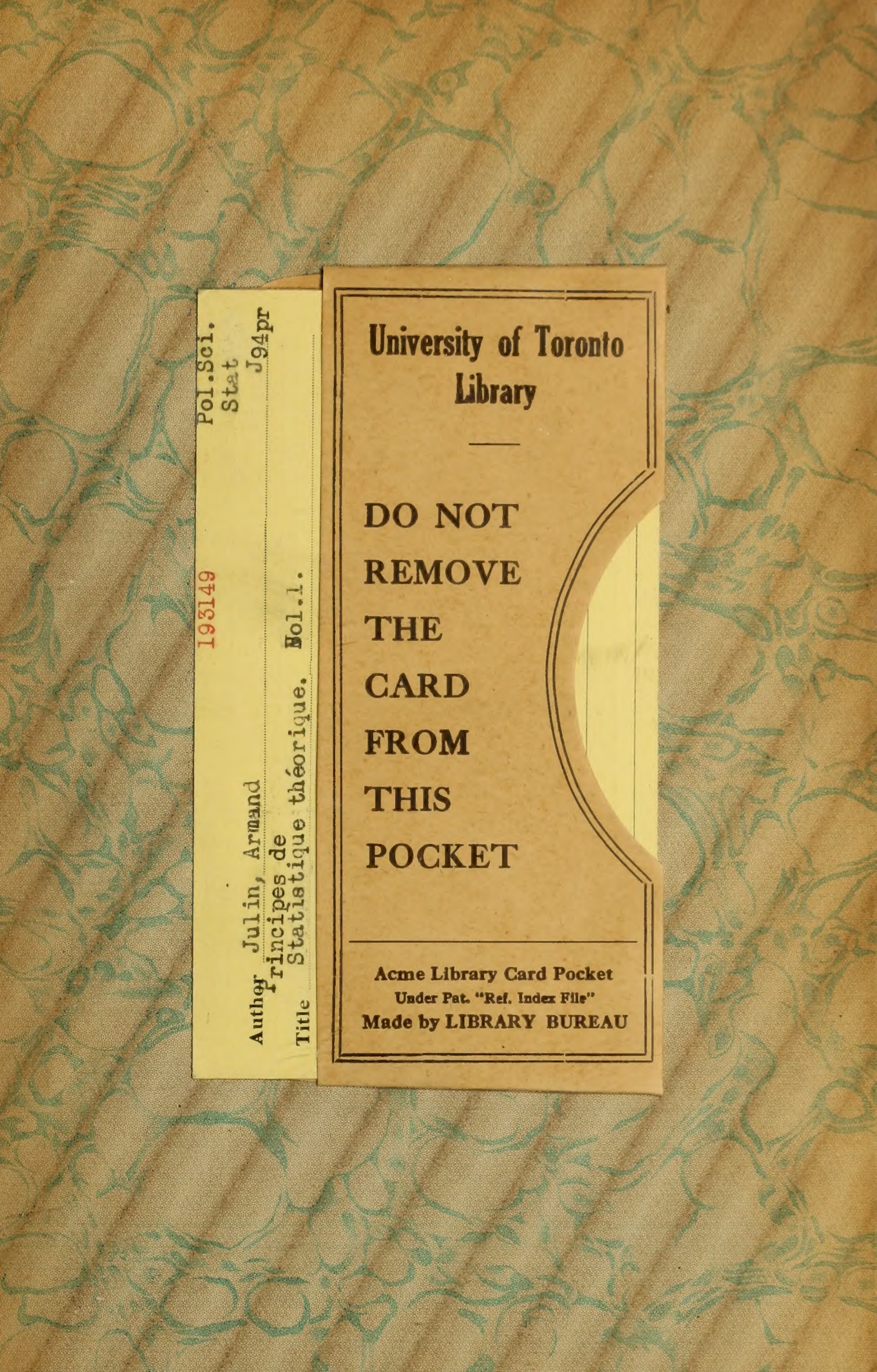
V. — Quelques déterminations d'erreurs probables.

TABLE ONOMASTIQUE

FIN

1952 — Soc. an. M. Weissenbruch, imp. du Roi, 49, rue du Poinçon, Bruxelles.





University of Toronto
Library
DO NOT
REMOVE
THE
CARD
FROM
THIS
POCKET
Acme Library Card Pocket
Under Pat. "Ref. Index File"
Made by LIBRARY BUREAU
Pol.Sci.
Stat
J94pr
193149
Author Julin, Armand
Title Principes de
Statistique théorique. Vol.1.

